



## **Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

### **Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

### **Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

### **Access Agreement**

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

# **An Investigation of and a Hybrid Recommender System for Evaluating Adaptive E-Learning Systems**

A thesis submitted to the

**University of Dublin, Trinity College**

for the degree of

**Doctor in Philosophy**

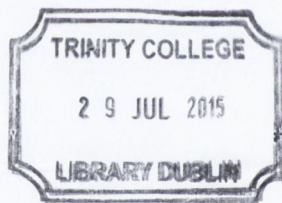


**Catherine Mulwa B.Sc, M.Sc**

Knowledge and Data Engineering Group,  
School of Computer Science and Statistics,  
Trinity College Dublin



Submitted May 2015

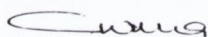


Thesis 10626

## Declaration

I, the undersigned, declare that this work has not been previously submitted as an exercise for a degree at this or any other University, and that, unless otherwise stated, it is entirely my own work.

I agree to submit this thesis in the University's open access institutional repository subject to Irish copyright legislation and trinity college library conditions of use and acknowledgement.



---

Catherine Mulwa  
May 2014

## Acknowledgement

I would like to express my sincere gratitude to my supervisors **Prof. Vincent Wade** and **Assistant. Prof Mary Sharp**, for their invaluable advice and support. I would also like to extend my gratitude to **Assistant. Prof. Seamus Lawless** and my colleagues who worked with me in the Knowledge and Data Engineering research Group (KDEG).

Special thanks to my examiners, **Prof. Mária Bielíková** and **Prof. Declan O'Sullivan**, and the chairperson of my defence, **Prof. Donal O'Mahony**, for their insightful comments.

My appreciation is extended to members of the Centre for Next Generation Localisation (CNGL) for partially fund this research and Trinity College Dublin (TCD) for the postgraduate studentship scholarship without the financial support I received, it would have been difficult to complete this research.

Many thanks also to my sister **Naomi**, my brothers **Shedrack** and **Daniel** for being so supportive and encouraging. Finally I would like to thank **Fr. Francis Caffrey C.S.Sp.** whose guidance and encouragement has helped me to grow as a Christian. My motivation comes from Jeremiah 29:11 which says: *“I know the plans I have for you. Plans to make you succeed and not fail.”*

This thesis is dedicated to my mum and to my three brothers (*Julius, Raphael and David*) who passed away recently during the course of this PhD. *“Mum, although it's many years since you left us, I still remember your final words to me; never give up kido, life is like a football game. You do not have much time to hit the target. To achieve your goals, you need to focus, and as long as you have your target in sight, strike without hesitation.”*

## Abstract

A key problem with research in the field of adaptive systems is the inconsistency of evaluation applied to such systems. A fact that is well established by expert evaluators is that adaptive systems cannot be evaluated as if they were non-adaptive. Several researchers acknowledge that evaluation of such systems is a difficult, demanding endeavour due to the complex nature of such systems. One major problem is the understanding of the adaptation mechanism of the system, what is improved by the adaptation, and what might have been the situation if a different kind of adaptation had occurred. Furthermore, when the evaluation of an adaptive system indicates a problem, such as user dissatisfaction and non-use of adaptive features, it is impossible to pinpoint the source of these problems, whether wrong user model, problems with the adaptation theory, wrong adaptation strategy, inappropriate method or evaluation techniques (methods, metrics, and criteria). It is important that evaluators of these systems use correct evaluation techniques.

This thesis investigates evaluations of adaptive E-Learning systems developed from 2000 to date and addresses the fact that it is difficult to identify the evaluation objective, the evaluation approach, and the range of evaluation choices. This evidence-based study examines what people have evaluated in adaptive systems and what evaluation techniques they used, and then maps those techniques to different evaluation approaches and techniques. Based on the results of these investigations, there is clear evidence that many design choices are being made during evaluations of adaptive E-Learning systems. For an expert evaluator this is tricky; for a novice evaluator it is much more difficult. Researchers need more advice on their evaluation options in order to attain their goal. They need support in their decision-making. To support these evaluators, the candidate has specified, designed and developed a web-based evaluation framework for supporting evaluators of adaptive systems (EFEx). In addition the candidate has designed and implemented a focused crawling system for evaluation studies of adaptive E-Learning systems.

The major contribution of this thesis is a novel hybrid (case-based and knowledge-based) recommendation service built on an evaluation educational dataset. A recommendation technology is used to enhance the appropriateness of suggestions for evaluation techniques for adaptive systems. A hybrid (case- study and user-centred) evaluation approach was taken to evaluate and validate the thesis. In addition a detailed analysis of the different aspects of the research is presented, outlining and addressing the identified challenges encountered by evaluators of adaptive systems.

# Table of Contents

<b>Declaration</b> .....	<b>i</b>
<b>Acknowledgement</b> .....	<b>ii</b>
<b>Abstract</b> .....	<b>iii</b>
<b>Table of Contents</b> .....	<b>iv</b>
<b>List of Figures</b> .....	<b>x</b>
<b>List of Tables</b> .....	<b>xi</b>
<b>v</b>	
<b>Abbreviations and Glossary</b> .....	<b>xvi</b>
<b>Chapter 1: Introduction</b> .....	<b>1</b>
1.1 Motivation.....	1
1.2 Research Question.....	2
1.3 Thesis Objectives .....	3
1.4 Research Methodology .....	4
1.5 Research Contribution.....	5
1.6 Thesis Overview .....	8
<b>Chapter 2: A Review of Evaluations of Adaptive E-Learning Systems</b> .....	<b>11</b>
2.1 Introduction.....	11
2.2 Summary and Critique of User-Centred Evaluations .....	12
2.2.1 Advantages: Software Evaluations .....	12
2.2.2 Advantages: Why User-Centred Evaluations are needed .....	13
2.2.3 Limitations: User-centred Evaluations .....	15
2.3 Summary and Critique of Adaptive Systems .....	17
2.3.1 Advantages: Why the Need for Adaptivity?.....	17
2.3.2 Adaptive Systems Variation Types & Internal models .....	22
2.3.3 Overview of Adaptive E-Learning Systems .....	24
2.3.3.1 Adaptive Hypermedia Systems.....	28
2.3.3.2 Adaptive Educational Hypermedia Systems .....	30
2.3.3.3 Intelligent Tutoring Systems.....	39
2.3.3.4 Adaptive Educational Game Systems .....	42
2.3.3.5 Adaptive Educational Hybrid Recommender Systems .....	46
2.3.3.6 Evaluation Frameworks for Supporting Evaluators.....	50
2.4 Comparison across Evaluation Approaches.....	51
2.4.1 User-Centred Evaluation Approach.....	52

2.4.2	Layered Evaluation Approach .....	53
2.4.3	Empirical Evaluation Approach.....	53
2.4.4	Utility-Based Evaluation Approach.....	53
2.4.5	Heuristic Evaluation Approach.....	54
2.4.6	Quality, Lifecycle & Combined Four-level and Six-level Approach . . .	55
2.4.7	Design Decisions: Critique of Evaluation Approaches.....	55
2.5	Tradeoffs between Techniques to Support User-Centred Evaluations.....	58
2.5.1	Summary and Critique .....	58
2.5.1.1	Evaluation Methods .....	61
2.5.1.2	Measurement Criteria .....	61
2.5.1.3	Evaluation Metrics .....	62
2.5.2	Design Decisions: Critique of Evaluation Techniques Used .....	63
2.6	Challenges, Problems, Issues and Limitations during Evaluations .....	66
2.6.1	Critique of Challenges .....	66
2.6.2	Critique of Problems and Issues .....	67
2.7	Conclusions .....	68
<b>Chapter 3: Overall Research Methodology &amp; EFEx and OSSES System Architecture</b>		
<b>Design</b> .....		<b>69</b>
3.1	Introduction .....	69
3.2	Influences from State of the Art .....	69
3.3	Overall Methodology .....	71
3.4	Overall Architecture .....	72
3.4.1	EFEx - Evaluation Framework for Supporting Evaluators of Adaptive Systems.....	72
3.4.2	OSSES - Focused Online Crawling Systems for Evaluation Studies.....	74
3.5	Conclusion.....	75
<b>Chapter 4: Evaluation Framework for Supporting Evaluators of Adaptive Systems</b>		
.....		<b>76</b>
4.1	Introduction .....	76
4.2	Objectives and Scope of EFEx Evaluation Framework .....	76
4.3	Architectural Design and Technological Design.....	77
4.3.1	Architecture Components and Capabilities .....	78
4.3.1.1	An Automated Novel Hybrid Automated Recommendation System .....	78
4.3.1.2	Personalized Search Sub-System for Evaluation Studies of Adaptive Systems .....	88
4.3.1.3	Taxonomy of Technical Terms .....	91
4.3.3	Design Testing.....	91



4.3.3.1	Experiment Objectives.....	92
4.3.3.2	Experiment Setup.....	92
4.3.3.3	Results and Findings.....	92
4.4	Prototype Implementation.....	95
4.4.1	Hybrid (Case-based & Knowledge-based) Recommender System.....	96
4.4.1.1	User Modelling: Authentication.....	98
4.4.1.2	Evaluation Approaches Used.....	99
4.4.1.3	Evaluation Methods.....	108
4.4.1.4	Measurement Criteria.....	112
4.4.1.5	Evaluation Metrics.....	113
4.4.1.6	Recommended Bundle (Method, Criteria & Metric).....	113
4.4.1.7	Dataset for the Recommender System.....	115
4.4.2	Personalized Search System.....	116
4.4.3	Taxonomy of Technical Terms.....	119
4.4.4	Prototype Testing.....	120
4.4.4.1	Evaluation Goal.....	120
4.4.4.2	Evaluation Set-up.....	121
4.4.4.3	Results and Findings.....	122
4.5	Conclusion.....	124
<b>Chapter 5:</b>	<b>Eliciting Knowledge-base for EFex.....</b>	<b>125</b>
5.1	Introduction.....	125
5.2	A User Study on Evaluations of Adaptive Systems Using an Evidence-based Approach.....	125
5.2.1	Evaluation Goal.....	125
5.2.2	Experiment Set-up.....	125
5.2.3	Results and Findings.....	126
5.2.3.1	Reported Adaptive Systems.....	126
5.2.3.2	Reported Evaluations Techniques (Methods, Criteria & Metrics).....	129
5.2.3.3	Internal Models of Adaptive Systems.....	134
5.2.3.4	Evaluations of the Internal Models of Adaptive Systems.....	135
5.3	Author Contribution: Educational Evaluation Dataset.....	137
5.4	Conclusion.....	141
<b>Chapter 6:</b>	<b>Evaluating EFEx – Recommendations Accuracy, Search Identification and Taxonomy Usability.....</b>	<b>143</b>
6.1	Introduction.....	143
6.1.1	Chapter Organization and Objective.....	145
6.2	A Hybrid Recommender System: Recommendation Appropriateness.....	145
6.2.1	Experiment for Novices.....	145

6.2.1.1	Experiment Objectives.....	146
6.2.1.2	Experiment Setup.....	147
6.2.1.3	Evaluation Results and Findings.....	149
6.2.2	Experiment for Experts.....	155
6.2.2.1	Experiment Objectives.....	155
6.2.2.2	Experiment Setup.....	155
6.2.2.3	Evaluation Results and Findings.....	156
6.2.2.4	Comparison-: Expert Results vs Recommendations produced by the Recommender System.....	166
6.3	Personalised Search System: Search Identification, User Satisfaction and Learnability.....	168
6.3.1	Experiment Objective .....	168
6.3.2	Experimental Setup.....	169
6.3.3	Results and Findings.....	170
6.4	Taxonomy of Technical Terms: Usability .....	177
6.4.1	Experiment Objective.....	177
6.4.2	Experiment Setup.....	177
6.4.3	Results and Findings (SUS Scores) .....	178
6.5	Conclusions .....	180
<b>Chapter 7:</b>	<b>Focused Online Crawling Systems for Evaluation Studies.....</b>	<b>182</b>
7.1	Introduction .....	182
7.2	Objectives and Scope of OSSES .....	182
7.3	Architecture and Technical Design .....	183
7.3.1	Architectural Design .....	183
7.3.2	Technical Design.....	185
7.4	Implementation.....	187
7.4.1	Administrator Component.....	187
7.4.1.1	Add System .....	187
7.4.1.2	RSS Feed Management .....	188
7.4.1.3	Crawl Management .....	188
7.4.1.4	Study Management.....	189
7.4.1.5	Evaluated System Management .....	190
7.4.2	Personalized Search Component .....	190
7.4.2.1	Personalized Search Components .....	191
7.4.2.2	Search Evaluated System .....	191
7.5	Prototype Testing .....	192

7.5.1 Evaluation Process .....	192
7.5.3 Results and Findings .....	192
7.6.1 System Usage .....	194
7.6 Conclusion .....	194
<b>8 Conclusion.....</b>	<b>196</b>
8.1 Research Question & Objectives Revisited .....	196
8.2 Contributions to the research field .....	205
8.3 Future Research Suggestions .....	209
<b>Bibliography.....</b>	<b>211</b>
<b>Appendixes .....</b>	<b>222</b>
Appendix A: List of Publications by the Author .....	222
Appendix B: Implementation of EEx .....	225
B1. Glossary of Ingredients of the Recommender Component .....	225
B2. Factors Considered When Recommending a Method .....	226
B3. Implementing a bundle (Method, Criteria & Metric) .....	227
B4. Calculating Total Score for Each Metric .....	231
B5. EEx Technical Design .....	232
4.3.2 Technical Design.....	232
Appendix C: Eliciting Knowledge-base for EEx.....	233
Appendix D: Evaluation Framework for End User Experience in Adaptive Systems (EEx.....)	236
D1. Post-Questionnaire: Design Testing of EEx .....	236
D2. Experiment 1-Hybrid (Case-based & Knowledge-based) Recommender System.....	238
D2.1 Expert Evaluators: Task-based Evaluation Experiment (Recommending Evaluation Techniques) .....	238
D3. Interacting with an Automated Hybrid (Case-based and Knowledge-based) Recommender System.....	247
D3.1 Novice Evaluators: Task-based Evaluation (Recommending Evaluation Techniques).....	247
D3.2 Novice Evaluators: Post Tasks Questions (SUS Questionnaire) .....	250
D3.3 Results: A Summary of SUS Scores by 31 Novice Evaluators .....	251
D4. Experiment 2 - A Personalised Sub- Search System.....	252
D4.1 Pre - Questionnaire .....	252
D4.2 Post -Tasks Questions (SUS Questionnaire).....	255
D4.3 Results: A Summary of the SUS Scores by the 33 Participants .....	256
D5. Experiment 3 - A Taxonomy of Technical Terms .....	257

D5.1	Identification of User Characteristics.....	257
D5.2	Post -Tasks Questions (SUS Questionnaire).....	258
D5.3	Results: A Summary of the SUS Scores by the 15 participants .....	259

# List of Figures

Figure 1- 1: Process of tackling the research question and objectives.....	4
Figure 2-1: 30 Years of highlights in development of desktop computing user evaluations, 1971–2001.....	13
Figure 2-2: Overview of usability challenges for user-adaptive systems (Jameson, 2009).....	16
Figure 2-3: S List of internal models used in analysed adaptive systems .....	24
Figure 2-4: Generalized architecture of an adaptive educational system .....	31
Figure 2-5: Generic layers in simplified example architecture of an educational AEH .....	31
Figure 2-6: Hierarchy of underlying factors of AEH (Mulwa, 2010 ) .....	32
Figure 3- 1: A Hybrid (evidence-based and user-Centred) research methodology .....	70
Figure 3- 2: High Level Overview of EFEx Framework Architectural Design.....	73
Figure 3- 3: High-level Architectural Design of OSSES Crawling Systems .....	75
Figure 4- 1: High Level Overview of EFEx Framework Architectural Design.....	78
Figure 4-2: Overview of the various components of the proposed recommendation service.....	79
Figure 4-3: Different aspects considered during the process of recommendation.....	82
Figure 4-4: Different aspects considered during Process of recommending an approach .....	86
Figure 4-5: User triggers ‘GetRecommended evaluation technique(s) .....	87
Figure 4-6: Process before a recommendation is produced.....	88
Figure 4-7: Architecture of the personalized search sub-system.....	89
Figure 4- 8: Taxonomy of technical terms architecture.....	91
Figure 4-9: Comparison rating of useful EFEx components .....	94
Figure 4-10: Automated EFEx evaluation framework home page .....	95
Figure 4-11: the CBR cycle, adapted from Choy et al., 2003 .....	97
Figure 4-12: User authentication before interacting with the recommender system .....	98
Figure 4-13: Expertise identification (novice or expert evaluator).....	99
Figure 4-14: Novice evaluator options: new system or existing system .....	100
Figure 4-15: Step 1 – System selection so that recommendations can start .....	101
Figure 4-16: Step 2 – Variation type of the system selected in step 1.....	102
Figure 4-17: Step 3 – Properties to select for focus during recommendations.....	102

Figure 4-18: Step 4 – Selecting of evaluation purpose .....	103
Figure 4-19: Step 5 – Selection of key aspects to focus on, based on evaluator’s choices.....	103
Figure 4-20: Step 6 – Selecting questions useful for the evaluation.....	104
Figure 4-21: Top 5 recommended approaches and explanations.....	107
Figure 4-22: Expert evaluator offered more options to customize recommendations ....	107
Figure 4-23: View recommended evaluation methods, criteria and metrics, or a combination of the techniques and recommended bundle .....	110
Figure 4-24: Ranked (Top 5) recommended evaluation methods.....	111
Figure 4-25: Recommended bundles (method, criteria and metric).....	115
Figure 4-26: Automated personalised search system UI .....	116
Figure 4-27: Inner metadata models of adaptive systems.....	117
Figure 4-28: Returned query results for navigation model .....	118
Figure 4-29: Results after query on APeLS system.....	119
Figure 4-30: Taxonomy of technical terms.....	120
Figure 4- 31: Experimental Process.....	121
Figure 4- 32: Comparison of useful features of EFEx .....	122
Figure 5- 1: Response to the question ‘Have you developed an adaptive system?.....	127
Figure 5- 2: Internal models used when developing an adaptive syste.....	134
Figure 6-1: Evaluation objectives of EFEx framework.....	144
Figure 6-2: Experimental Process for Novice Evaluators .....	148
Figure 6- 3: Identification of appropriate evaluation techniques .....	149
Figure 6-4: Percentile rank associated with SUS scores .....	151
Figure 6-5: Summary of SUS scores by novice evaluators .....	151
Figure 6-6: Recommender system (user satisfaction and learnability) .....	153
Figure 6-7: Variance in individual responses to questions, using standard deviation .....	154
Figure 6- 8: Experimental process for expert evaluators.....	156
Figure 6-9: User characteristics – identification of user experience and expertise of evaluators .....	157

Figure 6- 10: Explanations on recommended evaluation techniques .....	158
Figure 6- 11: Types of adaptive systems participants wished to focus on during evaluation.....	159
Figure 6- 12: Experts response on system characteristics.....	159
Figure 6- 13: Expert response on evaluation purpose .....	160
Figure 6- 14: Response percentage on questions .....	160
Figure 6- 15: Overall response to “what would it help you to improve in the system?” .....	161
Figure 6- 16: Recommended evaluation approaches .....	162
Figure 6- 17: Experimental setup of the personalised search system .....	169
Figure 6- 18: User characteristics of personalised .....	170
Figure 6- 19: Familiarity in using personalised search system.....	171
Figure 6- 20: Responses on return of relevant search results of internal models search system participants .....	172
Figure 6- 21: Responses on return of relevant search results of evaluations of adaptive systems .....	173
Figure 6- 22: Task 3 – Responses on return of relevant search results of general evaluation studies.....	173
Figure 6- 23: A summary of percentile scores of 33 participants of the personalised search system .....	174
Figure 6- 24: User satisfaction .....	176
Figure 6- 25: Frequency distribution of 33 participants – personalised search system .....	176
Figure 6- 26: Summary of percentile scores of the taxonomy .....	178
Figure 6- 27: User satisfaction and learnability.....	179

Figure 7- 1: High-level overview of various components of the focused crawling system..... 184

Figure 7- 2: Process of Crawling a published study ..... 185

Figure 7- 3: Relationship between an evaluation study and an evaluated system ..... 185

Figure 7- 4: Technological design of OSSES system ..... 186

Figure 7- 5: OSSES administrator component..... 187

Figure 7- 6: RSS feed management user interface..... 188

Figure 7- 7: Crawled feed and the URL link ..... 188

Figure 7- 8: Crawl management ..... 189

Figure 7- 9: Published study details..... 189

Figure 7- 10: Evaluated system detail..... 190

Figure 7- 11: Personalized search component ..... 190

Figure 7- 12: Search evaluated system ..... 191



## List of Tables

Table 2- 1: Summary of UCE evaluation techniques classification .....	15
Table 2- 2: Impact and advantages of adaptivity .....	18
Table 2- 3: Impact and advantage of adaptivity in E-Learning .....	20
Table 2- 4: Variation types (categories) of adaptive systems .....	22
Table 2- 5: Summary of E-Learning systems evaluation techniques.....	27
Table 2- 6: Summary of evaluations of adaptive educational hypermedia systems .....	34
Table 2- 7: Summary of intelligent tutoring systems, internal models and evaluation techniques.....	40
Table 2- 8: Summary of evaluations of adaptive educational games .....	45
Table 2- 9: Tradeoffs between recommendation techniques (Burke, 2002).....	47
Table 2- 10: Hybridization (combination) of recommendation methods .....	48
Table 2- 11: Design decisions made on which evaluation approaches to use and other approaches that could have been used .....	56
Table 2- 12: Tradeoffs between evaluation techniques (methods, criteria and metrics).....	59
Table 2- 13: Examples of design decisions on evaluation methods .....	63
Table 2- 14: Examples of design decisions on evaluation criteria.....	65
Table 4-1: Ingredients used by the recommendation algorithm for an evaluation technique.....	83
Table 4-2: Ingredients used by the recommendation algorithm for evaluation approach .....	85
Table 4-3: Identification of user characteristics (e.g. expertise in adaptive systems development).....	93
Table 4-4: Steps involved before recommending an approach.....	101

Table 4-5: Cases (Factors), Value, Weight and Reason considered when recommending an approach to be used.....	105
Table 4-6: Factors considered when recommending an evaluation method . . . . .	108
Table 4-7: Computing recommendations for an evaluation method . . . . .	109
Table 4- 8: Factors considered when recommending criteria. . . . .	112
Table 4-9: Factors considered when recommending an evaluation metric . . . . .	114
Table 4-10: Task-based questions on usefulness of EEx. . . . .	121
Table 6- 1: Evaluation Methods Experts would recommend . . . . .	163
Table 6- 2: Evaluation criteria experts would recommend . . . . .	164
Table 6- 3: Experts Recommended Metrics . . . . .	165
Table 6- 4: Expert -Response Recommended Bundles (Method, Criteria and Metric) .....	166
Table 6- 5: Results (Evaluation techniques) Produced by the Hybrid Recommender System.....	167

# Abbreviations and Glossary

<b>AEG</b>	Adaptive Educational Game
<b>AEHS</b>	Adaptive Educational Hypermedia Systems
<b>AEL</b>	Adaptive E-Learning Systems
<b>ALE</b>	Adaptive Learning Environment
<b>AHS</b>	Adaptive Hypermedia Systems
<b>DEG</b>	Digital Educational Game
<b>EAS</b>	Evaluation of adaptive systems
<b>EFEx</b>	Evaluation framework for supporting novice evaluators
<b>EQO</b>	European Quality Observatory
<b>ITs</b>	Intelligent Tutoring Systems
<b>ISO</b>	International Organization for Standardization
<b>MLE</b>	Managed Learning Environment
<b>TEL</b>	Technology-enhanced Learning
<b>TLCTS</b>	Tactical Language and Cultural Training System
<b>UCEA</b>	User-Centered evaluations approach (s)
<b>SUS</b>	System Usability Scale
<b>VLE</b>	Virtual Learning Environment

# **Chapter 1: Introduction**

## **1.1 Motivationchapter 1: Introduction**

### **1.1 Motivation**

The research field of adaptive E-Learning systems (AELS) has grown rapidly over the past 15 years and resulted in a range of terms, inner models, methodologies, and a plethora of new systems. AELS systems are becoming more popular as tools for user-driven access to information. This has led to the challenge of catering to a wide variety of users and generating appropriate information and user interfaces for them (Knutov et al., 2009), Mulwa C. et al., 2011). A key problem in the development of user interfaces is the inadequacy of traditional evaluation techniques to be used for the evaluation of adaptive user interfaces. In a research study conducted by (Paramythis et al., 2009), it was acknowledged that existing evaluation methods are only appropriate for assessing 'static' user interfaces. But not the way and extent to which dynamic adaptation facilities of the user interface affect interaction qualities, such as accessibility, usability, acceptability, etc. Furthermore, it is essential not only to evaluate the AELS but also to ensure that the evaluator uses the correct evaluation techniques since an incorrect technique can lead to wrong conclusions.

A key challenge emphasized by evaluators of adaptive systems is the difficulties due to the complex nature of such adaptive systems and the usability issues raised by catering for such diverse end users (Missier Del and Ricci, 2003, Lavie et al., 2005, Weibelzahl and Weber, 2002, Markham et al., 2003). One of the difficulties in evaluation of adaptation is to provide sufficient design feedback for the identification of problems and issues arising (Paramythis A., 2009). Evaluators are faced with difficulty in trying to understand the adaptation mechanism of the system; knowing what is improved by the adaptation, and what might have been the situation if a different kind of adaptation had occurred. In addition sometimes they face difficulties when defining the effectiveness of

adaptation. When users work with an adaptive system, it is difficult in principle to demonstrate what ‘might have been’ or what impact the system’s adaptive processes actually had on the end-user.

Furthermore major challenges faced by novice evaluators of such systems include usability issues such as difficulty in choosing the right evaluation approach and evaluation methods to use. In addition some of the typical goals of good usability principles for example predictability, transparency and uncontrollability may not be optimal for an adaptive system where the adaptive system wants to be able to change behaviour based on context of frequency of use and therefore can become less predictable (Jameson, 2009). Also traditional usability almost works against the notion of personalisation, however there is a benefit for and therefore is a tradeoff to be made.

Several researchers (Höök, 1997, Brusilovsky et al., 2004) highlight that evaluation is an important and challenging research issue in the area of adaptive learning systems (ALS) and adaptive systems. In fact, the lack of evaluation data, as well as the difficulty in their generalization, when available, and the resulting difficulty in the re-use of successful design practices, constitutes, among others, one of the main barriers for ALS to become mainstream technology. Furthermore, evaluation of these systems is a crucial and significant stage in their development (Jameson, 2009). These systems require some kind of evaluation due to their inherent usability problems at the interface and to ensure the correctness of adaptive solutions (Lawless et al., 2010, Tintarev and Masthoff, 2009).

## **1.2 Research Question**

This research work investigates current evaluation techniques used by evaluators of adaptive E-Learning systems and the tradeoffs between these techniques to support user-centered evaluations of such systems.

***RQ: “What are the techniques used and tradeoffs between the techniques which support user-centered evaluations of adaptive systems? Can a hybrid recommender system propose appropriate evaluation methods, criteria and metrics for individual adaptive systems and to what extent are these recommendations comparable to those of human expert recommendations.”***

In order to tackle the research question I divided it into two sub questions:

**Sub RQ1:** What are the techniques used and tradeoffs between the techniques which support user-centered evaluations of adaptive systems?

**Sub RQ2:** Can a hybrid recommender system propose appropriate evaluation methods criteria and metrics for individual adaptive systems and to what extent are these recommendations comparable to those of human expert recommendations.

### 1.3 Thesis Objectives

In order to address the research question, discussed in section 1.2, the following research objectives have been identified.

**Objective1** Investigate what are the capabilities and tradeoffs that user-centered evaluation (UCE) techniques can discover or estimate; through literature and survey.

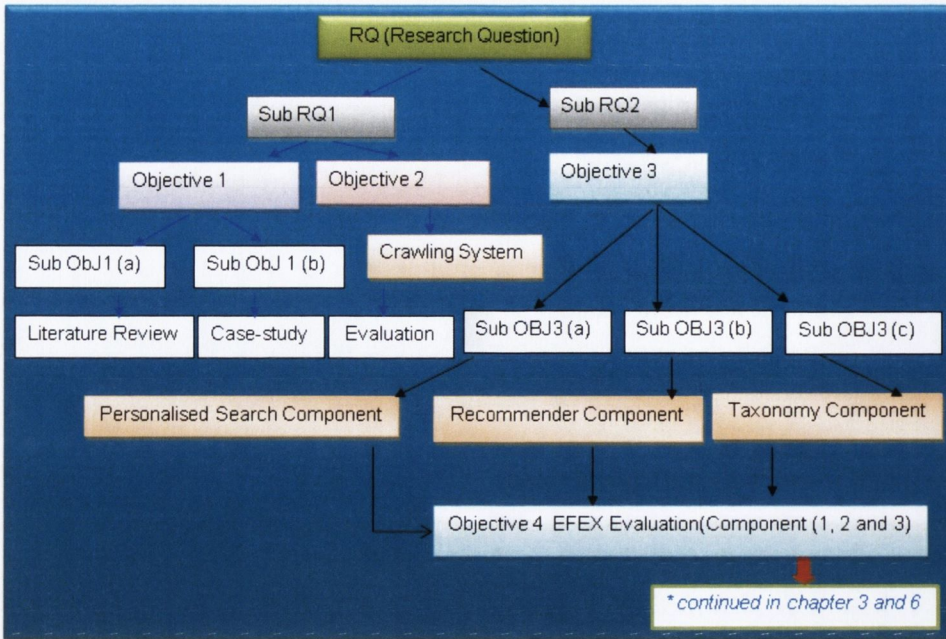
**Objective2** Design, develop and evaluate a focused crawling system for evaluation studies of adaptive systems.

**Objective3** Design and develop an evaluation framework for supporting novice and expert evaluators of adaptive systems (EFEx); which consists of three major components:

- (i) An automated hybrid (case-based and knowledge-based) system for recommending evaluation techniques,
- (ii) A personalised search component that allows users to find evaluation studies of adaptive systems and
- (iii) A taxonomy of technical terms for supporting the evaluation of adaptive systems.

**Objective 4** Evaluate the three components of the evaluation framework designed in Objective3

Figure 1-1 depicts the whole process of tackling the research question.



**Figure 1- 1: Process of tackling the research question and objectives**

## 1.4 Research Methodology

As stated in sections 1.1 and 1.2, this research focuses on providing a hybrid (case-based and knowledge-based) recommendation service for recommending evaluation techniques to assist different evaluators of adaptive systems. Furthermore it investigates the techniques used and tradeoffs between the techniques which support user-centered evaluations of adaptive systems. A summary and critique of these techniques and approaches in this field is presented in Chapter 2. Following this investigation, the approach applied in this research work consists of a hybrid of case-based study and user-centred approach.

The evaluation methods chosen for this research consists of: (i) structured interviews, (ii) task-based, real user studies, and (iii) online structured questionnaires. These were chosen because they are an effective method for measuring accuracy and appropriateness of recommendations, user satisfaction, learnability and usability of the evaluation framework. The methods also provide an effective technique for measuring the impact of techniques and technologies on user performance in real-world scenarios (He et al., 2008).

The analysis of the results in each evaluation cycle then enables further refinements to the developed architectures of the crawling system and the evaluation framework.

## 1.5 Research Contribution

This research work has resulted in one major and two minor contributions to the body of knowledge. It has also resulted in 17 peer-reviewed publications (in journals, conference papers, book chapters, workshops, demonstration events and posters).

The major scientific contribution of this thesis is a novel automated hybrid (case-based and knowledge-based) recommendation service for recommending evaluation techniques for adaptive systems. This service supports novice and expert evaluators to effectively and accurately identify appropriate evaluation techniques (or a combination of techniques) for such systems. This will encourage evaluations of such systems, especially AEL systems which are a focus of this research.

The first minor contribution consists of a personalized search system that supports novice evaluators in finding evaluation studies of: (i) internal models of adaptive systems; (ii) adaptive systems published from 2000 to 2013 and (iii) general evaluation studies of such systems. As part of this minor contribution on the search, I also had to develop this taxonomy of technical terms to help support the search.

These two contributions have resulted in the following peer-reviewed papers:

- Mulwa, C., and Wade, V. (2013). A Web-Based Evaluation Framework for Supporting Novice and Expert Evaluators of Adaptive E-Learning Systems. *Proceedings of International Conference on E-Technologies and Business on the Web (EBW2013)*, Society of Digital Information and Wireless Communication, pp. 62-67.
- Mulwa, C. Lawless, S., O’Keeffe, I., Sharp, M., and Wade, V. (2012). A Recommender Framework for the Evaluation of End User Experience in Adaptive Technology Enhanced Learning”. *International Journal of Technology Enhanced Learning*, pp. 67-84.



- Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2012). The Evaluation of Adaptive Technology Enhanced Learning Systems. *Proceedings of World Conference on E-Learning in Corporate, Government, and Healthcare and Higher Education (ELEARN)*, Association for the Advancement of Computing in Education (AACE), pp 744-753.
- Mulwa, C. Lawless, S., Sharp, M., and Wade, V. (2011). A Web based Framework for the Evaluation of End User Experience in Adaptive and Personalised E-Learning Systems. *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE Computer Society, vol. 3, pp 351-356
- Mulwa, C., Lawless, S., Ghorab, M.R, O'Donnell E, Sharp, M., Wade, V. (2011). A Framework for the Evaluation of Adaptive IR Systems through Implicit Recommendation. *Proceedings of the 19th international conference on Conceptual structures for discovering knowledge*. Springer Berlin Heidelberg, pp 366-374.
- Mulwa, C., Lawless, S., Sharp, M., Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems a (Demonstration Paper). *Proceedings of User Modeling, Adaptation and Personalization Conference, (UMAP 2011)*.
- Mulwa, C., Lawless, S., Sharp, M., Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems, a (Poster Paper). *Proceedings of User Modeling, Adaptation and Personalization Conference (UMAP 2011)*.
- Mulwa, C., Longo, L., Lawless, S., Sharp, M., and Wade, V. (2011). An online framework for supporting the evaluation of personalised information retrieval systems. *Proceedings of the 6th international conference on Ubiquitous and Collaborative Computing*. British Computer Society, pp 75-85.
- Mulwa, C., Li, W., Lawless, S., and Jones, G. (2010). A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction. *In Proceedings of the Workshop on Simulation of Interaction: Automated Evaluation of Interactive IR at SIGIR 2010*.

The second minor contribution is an educational evaluation dataset for adaptive systems. The dataset, based on peer-reviewed evaluation cases, is a characterized, structured and interlinked list of evaluation approaches, methods, metrics and measurement criteria extracted from over 350 papers in the literature on adaptive systems. Thus, rather than being a large dataset based on many users' behaviour, it is based on a smaller dataset that has been quality-reviewed. Moreover, the dataset can grow overtime as the framework itself provides a mechanism for published researchers to add their evaluation cases to it; thus the dataset is already a very valuable aid to adaptive E-Learning evaluation choices. Running multiple AEL systems over the dataset could provide a means of comparing recommender systems' accuracy. Using a combination of web crawling services, evaluation studies published from 2000 to 2013 were manually sliced to extract such a dataset.

This contribution has resulted in the following peer-reviewed paper:

- Mulwa, C., Lawless, Sharp, M. and Wade, V. (2011). The Evaluation of Adaptive and User Adaptive Systems: A Review. *International Journal of Knowledge and Web Intelligence (IJKWI)*, pp 138-156.
- Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I., and Wade, V. (2010). Adaptive educational hypermedia systems in technology enhanced learning: a literature review. *Proceedings of the 2010 ACM Conference on Information Technology Education*, pp 73-84.

A final output of this thesis is a focused crawling system for evaluation studies of adaptive systems (OSSES) that had to be built in order to assist researchers who are in the earlier stages of research when conducting literature review. This contribution has resulted in one peer-reviewed conference paper:

- Mulwa, C., Lawless, S., Sharp, M. and Wade, V. (2010) "OSSES: An Online System for Studies on Evaluation of Systems", *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Association for the Advancement of Computing in Education (AACE), pp 3210-3219.

Furthermore, as a result of investigating existing literature in the fields of evaluations of adaptive hypermedia, recommender systems and E-Learning, the following papers were published:

- Mulwa, C., McDonald, H., O'Keeffe, I., Lewis, D., He, Y., and Wade, V. (2012). The Maturity Model: A Novel Way of Evaluating Centre for Next Generation Localisation Demonstrator Systems Based on Industrial Impact, Scientific Collaboration and Interoperability. *Proceedings of World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE), pp 1340-1349.
- Mulwa, C., Lawless, S., Sharp, M., Sanchez, I. A., and Wade, V. (2010). Adaptive Educational Hypermedia Systems in Technology Enhanced Learning: A Literature Review. *Proceedings of the 2010 ACM conference on Information Technology Education*, pp 73-84.
- Lawless, S., O'Connor, A. and Mulwa, C. (2010). A Proposal for the Evaluation of Adaptive Personalised Information Retrieval", *Proceedings of the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation held in conjunction with ECIR2010 European Conference on Information Retrieval*.

A full list of these papers, posters and presentations resulting from this thesis can be found in *Appendix A*.

## **1.6 Thesis Overview**

This thesis proposes an innovative recommendation approach for supporting novice evaluations of adaptive systems, more specifically focusing on AEL systems. The thesis describes the design and implementation of three systems: i) A hybrid recommender system for supporting novice and expert evaluators of adaptive systems, ii) a personalised search system for evaluation studies of adaptive systems and iii) an online crawling system for evaluation studies. Finally it describes the design and implementation of a taxonomy of technical terms of adaptive systems

The remainder of this thesis is organized as follows. Chapter 2 presents a critical review of evaluations of adaptive E-Learning Systems (AELS). In particular it focuses on user-centred evaluations and comparisons across different evaluation techniques and approaches used. In addition the chapter discusses the tradeoffs between evaluation techniques to support user-centred evaluations. It also presents current challenges, problems, issues and limitations encountered by different evaluators during evaluations of such systems.

Motivated by the identified gaps and the overall findings of chapter 2, chapter 3 discusses the overall research methodology and the proposed architecture of an evaluation framework for supporting novice evaluators of adaptive systems (EFEx). It further describes the overall architecture of a focused crawling system for evaluation studies of adaptive systems.

Based on these influences of the state of art, Chapter 4 describes the design and implementation of the evaluation framework for supporting novice evaluators of adaptive systems (EFEx) which consists of three main components: (i) a hybrid recommendation system, (ii) a personalised search system and (iii) a taxonomy of technical terms of evaluations of adaptive systems. The chapter further discusses testing and validation done during design and implementation. In addition the results and findings of these evaluations are also presented.

Chapter 5 discusses eliciting of knowledge-base for EFEx. This involved conducting a real life user study of evaluations of adaptive systems developed from 2000 to 2012. The results and findings of this study are also presented. The chapter further presents an evaluation educational dataset resulting from this study. This dataset is used to populate the rules of the EFEx framework presented in chapter 4.

Chapter 6 discusses the evaluation experiments conducted to validate accuracy, usability and learnability of the developed components in chapter 4. During these evaluations a hybrid (of interview-based, task-based and real life user trials) of evaluation methodologies were used. In addition it presents the evaluation results and findings.

Chapter 7 further describes the design and implementation of a crawling service for evaluation studies (OSSES). Also the results of evaluations are presented. This tool

supports new researchers especially those conducting literature reviews of evaluations of adaptive.

Finally chapter 8 concludes this thesis with a summary of the findings and contributions of this research. The chapter also suggests future research directions in the field of evaluation of adaptive E-Learning systems.

# Chapter 2: A Review of Evaluations of Adaptive E-Learning Systems

## 2.1 Introduction

This chapter presents a summary and critical review of the state of the art of evaluations of adaptive E-Learning (AEL) systems developed from 2000 to 2013. An evidence-based approach is taken to investigate the evaluation techniques (methods, criteria and techniques) used and trade-offs between these techniques to support user-centred evaluations of AEL systems. All the classifications<sup>1</sup> presented in this chapter are created by the author.

Section 2.2 briefly summarizes and critiques the benefits and limitations of user-centred evaluation approach. Section 2.3 presents an overview and Critique of Adaptive Systems. Specifically focus on different variation types and the internal models of such systems. Furthermore the section critically reviews adaptive E-Learning Systems. Majority of AEL systems are adaptive hypermedia (AH) systems. AH systems stem from the information access paradigm of searching by browsing, where different groups of users generally have less precise information needs and therefore need to browse and explore pages. Finally the section presents a summary of evaluation frameworks for supporting evaluators of adaptive systems in general.

Evaluators of adaptive E-Learning systems have used different evaluation approaches. Section 2.4 presents a comparison across of evaluation approaches. Such evaluators have made different design decisions during evaluations. A critique of these decisions on which approach to uses is also presented. Section 2.5 introduces tradeoffs between evaluation techniques to support user-centred evaluations, design decisions and a critique of evaluation techniques used by different evaluators of AEL systems. Furthermore section 2.6 presents an overview of challenges, problems and issues encountered by novice and expert evaluators of such systems. Finally section 2.7 concludes the chapter.

---

<sup>1</sup> All classifications presented in Section 2.32 to Section 2.6 are produced by the author.

## **2.2 Summary and Critique of User-Centred Evaluations**

### **2.2.1 Advantages: Software Evaluations**

Evaluations are important tools of software quality assurance. Currently, there are several definitions of evaluation. Worthen et al. (1997) define evaluation as the “identification, clarification, and application of defensible criteria to determine an evaluation object’s value, quality, utility, effectiveness, or significance in relation to those criteria” (Worthen et al., 1997). Another researcher has defined evaluation as “the process of examining the product, system components, or design, to determine its usability, functionality and acceptability, which is measured in terms of a number of criteria essential for any software development project” (Weibelzahl, 2003).

Farooq (2008) emphasizes that a typical software quality program involves: (i) establishment, implementation and control of requirements, (ii) establishment and control of methodology and procedures, and (iii) software quality evaluation (Farooq, 2008). The software quality evaluation component is aimed at evaluating products both in process and at completion time, and methodologies for appropriateness and technical adequacies have been used. It is important to evaluate all software products, and to ensure that the evaluation uses the correct method (Brusilovsky, 2004).

System evaluation places an emphasis on the comparison of the presented system with established criteria proposed by other researchers or other related systems. The process applied involves the systematic determination of merit, worth and significance. In software development, evaluations are used to determine the quality and feasibility of preliminary products such as mock-ups and prototypes as well as of the final system. Evaluation also has the advantage of providing useful feedback to the developer for subsequent redesigns.

A key significance of evaluation is that its results and findings can offer valuable insights about the real behaviour and preferences of users. They can demonstrate that a certain adaptation technique actually works, i.e., that it is accurate, effective and efficient. Evaluation studies are an important means to convince evaluators, customers or investors of the usefulness and feasibility of a system. Furthermore, evaluations are important for scientific advancement as they offer a way to compare different approaches and techniques.

It is important to evaluate all software products and to ensure that the evaluation uses the correct evaluation techniques (method, criteria and metrics) (Brusilovsky, 2004d). This is emphasized more from our earlier research work on user-centred evaluation (UCE) of adaptive systems (Mulwa C. et al., 2010, Lawless et al., 2010).

### 2.2.2 Advantages: Why User-Centred Evaluations are needed

The International Organization for Standardization (ISO) defines usability of a product as “the extent to which the product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use”. Key attributes of usability are: learnability, efficiency, memorability, errors, and user satisfaction. The discipline of usability engineering provides structured methods for achieving usability in user interface design during product development; usability evaluation is part of this process. Figure 2-1 shows a timeline for usability evaluations in the last 30 years (Scholtz, 2004). Users were first used as the source of usability feedback but models have also been used for over 20 years. Expert feedback was developed in heuristic reviews and cognitive walkthroughs and has been used since the early 1990s. All three methods rely on usability engineers or usability professionals to design, conduct, analyze and report on the evaluations.

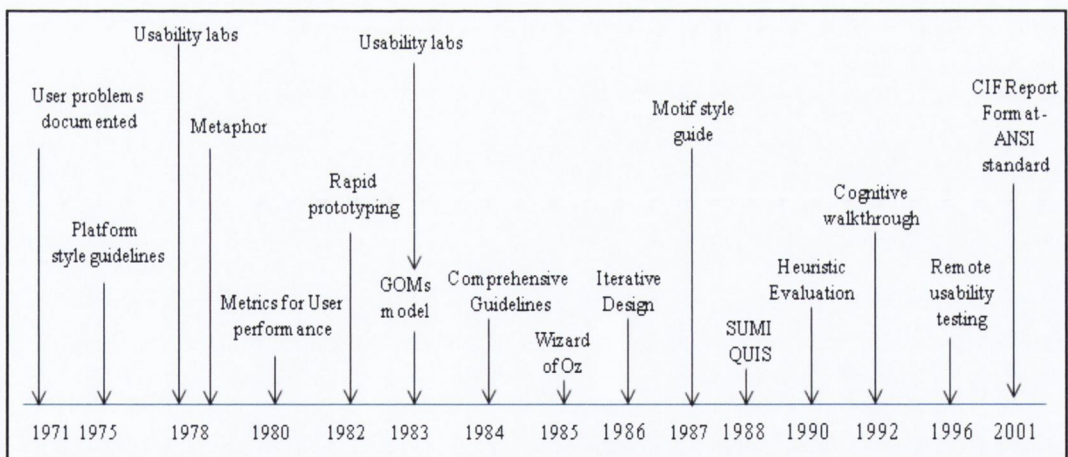


Figure 2-1: 30 Years of highlights in development of desktop computing user evaluations, 1971–2001

The main advantage of user-centred evaluation (UCE) is the involvement of users. Results are based on actually seeing what aspects of the user interface cause problems for representative users. In order to accomplish UCE evaluations, both novice and expert



evaluators need to identify representative users and representative tasks, and also implement a procedure for capturing the problems and issues that users' have when trying to apply a particular software product.

Furthermore, UCEs are very important during the design, testing and implementation cycle of system implementation; in most cases two types of evaluations are carried out. Formative evaluations are used to obtain information used in design, while summative evaluations are usability evaluations that document the effectiveness, efficiency and user satisfaction of a software product at the end of the implementation cycle. Representative users are recruited during both types of evaluations; some evaluation techniques for collecting information are used, and ways of disseminating the results of the evaluation to members of the software development team are needed.

UCEs serve several goals; for example, verifying the quality of an AEL system, detecting problems in the system functionality or interface, and 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 a system or help developers to decide which version of a system to release. The benefits of the user-centred approach are time and cost savings, ensuring the completeness of system functionality, minimizing required repair efforts, and improving user satisfaction (Nielsen, 1993).

The User-Centred Evaluation (UCE) Approach (Lawless et al., 2010) has proved to be useful in verifying the quality of AEL systems; detecting problems in the system functionality and user interface, and supporting adaptivity decisions (De Jong and Schellens, 1997). A lot of research has been conducted since the early 1970s (see Figure 2-1). Recently, adaptive system researchers have attempted to classify UCE evaluations into several categories (see Table 2.1). Three phases of UCE evaluation are identified (i.e. requirement, preliminary and final phase) as well as the evaluation methods established. It is evident that there is tradeoff between evaluation methods, variables assessed, and evaluation metrics across the different phases during evaluations. Table 2-1 presents a summary of evaluation techniques used from 2000 to date and the tradeoffs between different evaluation techniques.<sup>2</sup>

---

<sup>2</sup> The exchange of one thing for another of more or less equal value, especially to effect a compromise

**Table 2- 1: Summary of UCE evaluation techniques classification**

<b>User-Centred Evaluation</b>				
<b>Classification</b>	<b>Phase of evaluation</b>	<b>Evaluation method/instruments</b>	<b>Variables frequently assessed</b>	<b>References</b>
Observation & monitoring usage	Requirement	User observation, systematic observation, verbal protocol & cultural probes	Usability, user behaviour	(Santos Jr et al., 2005, Gena, 2005, Van Velsen et al., 2008)
Collection of users' opinions	Preliminary	Interviews, questionnaires (online, post-test, pre-test, pre-post-test, focus group & discussion group)	Usability, perceived usefulness, intention to use, trust & privacy issues, appropriateness of adaptation	(Gena, 2005, Van Velsen et al., 2008)
Formative evaluations	Preliminary	Wizard of Oz simulation, scenario-based design & prototypes	Early prototype evaluations, evaluations before implementation	(Masthoff, 2006)
Predictive evaluation	Hybrid (requirement & preliminary)	Heuristic evaluation, expert review, parallel design, cognitive walkthroughs & social-technical models	Usability of interface adaptation, user domain & interface knowledge, privacy transparency, appropriateness	(Van Velsen et al., 2008, Gena, 2005)
Experiments & tests	Final	Usability testing, experimental evaluation	Interface (& content) adaptation	(Van Velsen et al., 2008, Gena, 2005)

### **2.2.3 Limitations: User-centred Evaluations**

The downside of UCE is that user evaluations are expensive and time-consuming (Scholtz, 2004). Finding and scheduling an appropriate number of representative users for each user type is difficult. Laboratory and usability engineering resources are needed to conduct the evaluations and analyze the results. There are also issues involved as to the 'realism' of the evaluation. Have the correct tasks been selected? How will the product work in real work environments? Beta-testing and user feedback after installation are used to gather data about usability aspects of the product in the actual context of use.

Jameson (2009) identified a different perspective on potential problems emanating from the introduction of adaptivity in a system. The researcher identified usability challenges and problems related to adaptivity, their typical properties, and possible preventive and compensatory measures that can be employed to address these challenges. The challenges are expressed as usability goals to be met by evaluators (see Figure 2-3). For example, evaluators of adaptive systems encounter usability problems such as the following: (i) usability goals correspond to several desirable properties of interactive systems; (ii) predictability, transparency, controllability and unobtrusiveness correspond to general usability principles; (iii) maintenance of privacy and breadth of experience are relevant to adaptive and personalized E-Learning systems, and (iv) the column of typical properties lists examples of frequently encountered properties of these systems. Each has the potential to cause difficulties with respect to one or more of the usability goals: (v) the preventive measures aim is to ensure that a property is not present in such a manner that it would cause problems, and (vi) the compensatory measures goal is to ensure that, in some other way, the goals and objectives are achieved despite the threats created by the properties challenges (Jameson, 2009).

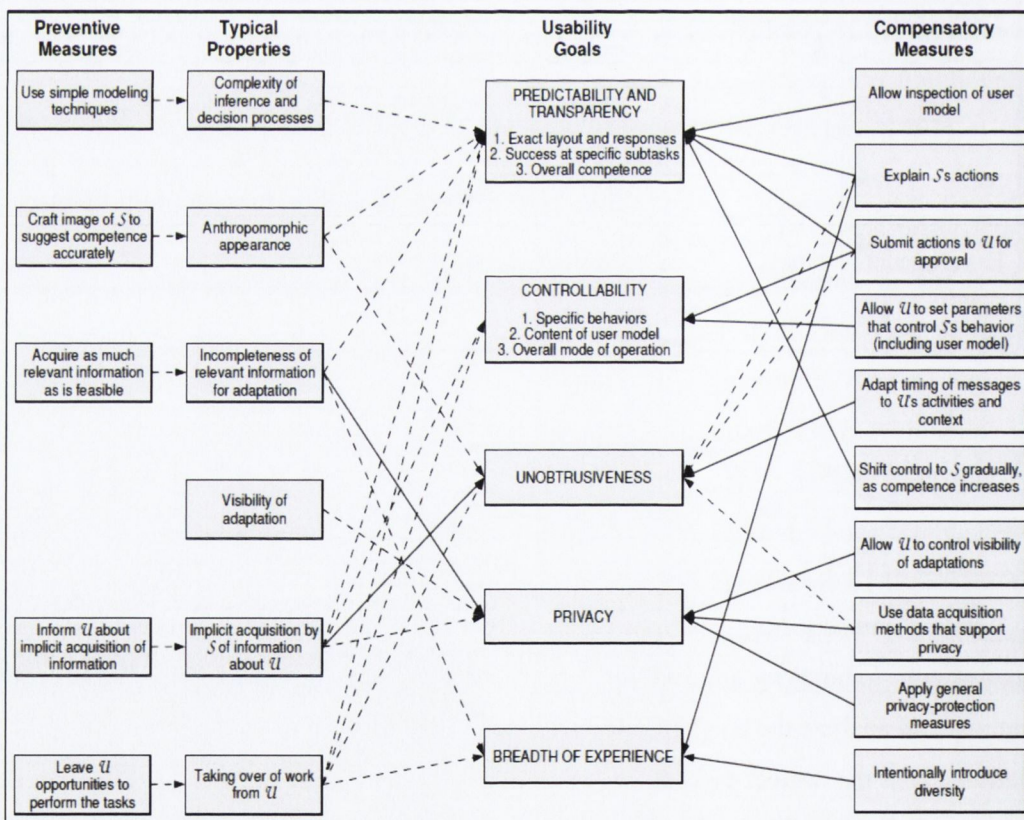


Figure 2-2: Overview of usability challenges for user-adaptive systems (Jameson, 2009)

These systems require some kind of evaluation due to their inherent usability problems at the interface and to ensure the correctness of adaptive solutions. The main focus in this research work has been advocating the importance of UCE evaluations, especially when evaluating adaptive systems due to their complex nature. The next section presents a critique of adaptive systems developed in the past 15 years.

## **2.3 Summary and Critique of Adaptive Systems**

The research field of adaptive systems has been growing rapidly over the last 15 years, and resulted in a range of 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 wider variety of users in differing environments and to user trust issues.

The evaluation of the adaptive system is of utmost importance and should be conducted using the correct evaluation techniques. It is important to not only evaluate but also ensure that the evaluation uses the correct methods (Brusilovsky et al., 2004) and the correct criteria and metrics (Mulwa C. et al., 2011) since an incorrect method, criteria or metric can lead to wrong conclusions (Gena and Weibelzahl, 2007, De Jong and Schellens, 1997). The evaluation of these systems is a fundamental stage in their development and it should become common practice. As the application of these systems moves from the research lab to real field usage, the evaluation of the real user-system interaction becomes fundamental. Since most of the time the exploitation of such techniques makes the system more complex, it is necessary to evaluate whether the adaptivity really improves the system, and whether the user really prefers the adaptive version compared to the non-adaptive.

### **2.3.1 Advantages: Why the Need for Adaptivity?**

Currently there are several definitions of adaptive systems. According to (Oppermann, 1994), a system is called adaptive if it is able to change its characteristics automatically according to users' need. On the other hand, adaptive systems consider the way the user interacts with the system and modify the interface presentation or the system behaviour accordingly (Weibelzahl, 2003). Sometimes adaptivity is confused with adaptability. A system is considered adaptable if it provides the user with tools that make it possible to change the system characteristics (Oppermann, 1994). Jameson (2001) summarized both

adaptivity and adaptability in the term personalization. Adaptive systems adapt their behaviour to the user and/or the user's context.

A key challenge for developers of adaptive systems is; understanding what is improved through adaptivity. Studies in this field rarely report the impact of adaptivity. This lack of research motivated us to conduct a real-life, evidence-based study,<sup>3</sup> where users were asked, if they had developed an adaptive system between 2000 and 2011; if they responded yes, we then asked them “what was improved by adaptivity?” (Mulwa C. et al., 2012). This study was aimed at three communities (user modelling, adaptive hypermedia and adaptive recommender), and there were 96 participants in the study. Users reported 77 systems. In response to the question of adaptivity, only one developer reported that “nothing was improved by adaptivity”. During the analysis of the results (refer chapter 5), what was perceived to have been improved by adaptivity was categorized into eight groups (personalization, technology-enhanced learning, user satisfaction, results output, time, recommendations, adaptation). Table 2-2 presents these results exactly the way users responded, as well as the named system.

**Table 2- 2: Impact and advantages of adaptivity**

<b>Adaptivity</b>		
<b>Category</b>	<b>What was improved</b>	<b>System name</b>
Adaptive & adaptation	<ul style="list-style-type: none"> <li>· Adaptive testing. The main advantages are: - Reduce the number of questions to be posed (reduce required time) - Alternatively, keeping the same number of questions, increase test reliability</li> <li>· Adaptation of: - background of student's (course and discipline)</li> <li>· navigation preferences - knowledge level</li> </ul>	SIETTE (1998-2011)  AdaptWeb (2001)
Personalization – because it tailors what is presented to that individual	<ul style="list-style-type: none"> <li>· Personalization &amp; personalized teaching</li> <li>· Personalized books were generated based on student model.</li> <li>· Yes the personalized search results were more relevant to users. The number of failed searches was reduced, the time to complete</li> </ul>	NewsAtHand (2009), PERSONF, ActiveMath (2000)  aLFanet (2002-2005),

<sup>3</sup> <http://www.surveymonkey.com/s/Q2DSDF8>

	<p>searches reduced etc.</p> <ul style="list-style-type: none"> <li>· Adapted instructional design and personalized guidance to students in terms of recommendations.</li> <li>· The personalized search results were more relevant to users. The number of failed searches was reduced, the time to complete searches reduced etc.</li> </ul>	I-SPY (2003)
Recommendations & explanations	<ul style="list-style-type: none"> <li>· The users were more satisfied with the system, but only when it explained the adaptation to the user.</li> <li>· Recommendations</li> <li>· Explanatory facility</li> <li>· Music video recommendations.</li> </ul>	ExDis (2011) Personalized explanations Rec. (2006-2009), APIL (2010)
Time	<ul style="list-style-type: none"> <li>· The length of the test and report on student knowledge.</li> <li>· Reduce required time alternatively by keeping the same number of questions.</li> <li>· Increase test reliability.</li> </ul>	GATA (2010)
Results output	<ul style="list-style-type: none"> <li>· The ranking of search results obtained from search engines was adapted to user needs.</li> <li>· The quality of the results and explanatory facility.</li> <li>· Search engine results (i.e. the system was particularly focused on query disambiguation).</li> <li>· Accuracy of recommendations.</li> <li>· Training plan, scenario, setting; intensity and amount of exercise required; recommendations.</li> </ul>	Search Behaviour-Driven Training for Result Re-ranking, Bifrost (2009-2010) & PIA EMSAVE, Monster and Gold, MOPET, GeoKaos+Flareqoor
User satisfaction	<ul style="list-style-type: none"> <li>· User efficiency, effectiveness and satisfaction</li> </ul>	Adaptive Information Retrieval and Composition System (2010)

Most of the reported systems belonged to AEL systems. The evidence is clear from users of these three communities that adaptivity improves the quality of such systems, which

leads to increased learnability and user satisfaction. Table 2-3 presents a summary of users' response to what was improved by adaptivity and the name of the system.

**Table 2- 3: Impact and advantage of adaptivity in E-Learning**

<b>Adaptivity</b>		
<b>Category</b>	<b>What was improved</b>	<b>System name</b>
Adaptive E-Learning – because they lead to better learning through presentation of most relevant learning material	<ul style="list-style-type: none"> <li>· Sequencing of the learning materials navigation support selection of tasks to work at feedback and selection of tasks based on students' current knowledge and preferences.</li> <li>· The aim was to provide students with tailored education in the way of choosing an appropriate level of difficulty. Additionally the system provides a course generator for different learning scenarios, such as preparing for an exam. Adaptivity is also included in the way of providing colour-flag feedback and the availability of hints, increasingly offering more information about the correct solution or the path towards the solution.</li> <li>· Privacy protection &amp; Tailoring content to specific users.</li> <li>· Content that is presented to users.</li> <li>· Awareness and learning support.</li> <li>· Adaptation of: - background of student's (course and discipline) - navigation preferences - knowledge level.</li> <li>· Presentation and interaction.</li> <li>· The content and the navigational guidance provided to students depending on personal features, actions and current context (device, time and physical location).</li> <li>· The recommendations about what to learn next.</li> <li>· Feedback and selection of tasks based on students' current knowledge and preferences.</li> <li>· The learning outcome and improved English learning at early ages (3 to 6 years old).</li> <li>· Students solved a science problem-solving scenario more quickly, and received pedagogical supports that were tailored to their curricular knowledge and</li> </ul>	<p>NetCoach (2000), SPORAS (2004)</p> <p>Friend Finder, Late-o-Meter, Contextual Display, 2010</p> <p>Adaptive extension to an E-Learning platform (2008-2010)</p> <p>JTS, ViSMod, The Learning Game, BELLA/English-Math ABLE (EM-ABLE Radiotube.</p>

Table 2-3 Continued

Adaptivity		
Category	What was improved	System name
	<ul style="list-style-type: none"> <li>· Problem-solving behaviours.</li> <li>· Studying behaviour; engagement with lifelong learning; mathematical generalization; theory-aware learning design.</li> <li>· The ability of users with a tremor disorder to separate deliberate motions from involuntary motions.</li> <li>· Navigational abilities of the robot, ability to escape from traps, speed with which robot could complete the task.</li> <li>· Transferability between different robotic platforms.</li> <li>· The sequence of materials shown to students.</li> </ul>	CoMoLE (2007-2008) OWL (2000)
	<ul style="list-style-type: none"> <li>· Students can see the model of their level of understanding in a range of topics, and make informed decisions about their learning. They can also use this information as a basis for peer collaboration.</li> <li>· Useful adaptation, in the form of link annotation/hiding and the conditional inclusion of fragments.</li> <li>· Some personalized teaching (e.g. matching the information to the learner (e.g. SASY's demonstrators), reducing the amount of information displayed (e.g. Locator) and interpreting information about the user differently (e.g. Locator).</li> <li>· Student's intelligent skills.</li> <li>· Studying behaviour; engagement with lifelong learning; mathematical generalization; theory-aware learning design</li> </ul>	SHAIEX Crystal Island (2008-2011) Idiotypic control network for a navigating mobile robot (2006-2009) OLMlets Inspire (2000), MyPlan (2008), Migen (2010), Learning Designer- (2011)

The results and findings of this study produce clear evidence that adaptivity is an important feature of an adaptive system. Furthermore, evaluators should ensure that the correct evaluation techniques are used, especially when evaluating AEL systems (discussed in section 2.3.3). A major significance of adaptive systems is that they are used in many domains to solve different tasks. Jameson (2001) categorized some of these adaptive functions into: (i) help the user to find information, (ii) tailor information to user, (iii) recommend products, (iv) help with routine tasks, and (v) adapt an interface and support learning (Jameson, 2001).



This research work categorizes adaptive systems into 14 variation types (categories) based on the tasks the systems perform. The next section briefly discusses these categories.

### 2.3.2 Adaptive Systems Variation Types & Internal models

Adaptive systems have been applied in several domains. Developers of such systems categorize them based on the tasks the system performs and the field in which it is used. A study was conducted in Chapter 6 which aimed at identifying different variation types (categories) and the internal models used during implementation of such a system. Table 2-4 presents a summary of adaptive systems reported in the study and the variation types (categories) to which each belongs. Majority of these systems belonged to adaptive educational systems.

**Table 2- 4: Variation types (categories) of adaptive systems**

<b>Evaluations of Adaptive Systems Developed between 2000 and 2011</b>	
<b>Variation types (categories)</b>	<b>System name</b>
Adaptive Educational Hypermedia Systems	UNITE (2008 – 2011), Dashboard at KiWi Framework, PERSONF, Locator, SASY, Personis, VLUM/SIV, aLFanet, 2002-2005, EMSAVE (2009-), OLMlets (2006), SPORAS (2004), SIETTE (1998-2011), Inspire (2000), MyPlan(2008), Migen (2010), CoMoLE (2007-2008), GATA (2010), SQL-Tutor, EER-Tutor (2003), ERM-Tutor (2005), J-Latte (2006), AdaptWeb (2001), MEDEA (2006), NetCoach (2000), Activemath (2000), ActiveMath (2000-2011).
Adaptive Learning Systems	Peer Finder, SHAIEX
Adaptive Information Retrieval Systems	I-SPY (2003), MovieLens (2000-2006), Radiotube.tv, SuggestBot (Wikipedia, 2005-), Search Behaviour-Driven Training for Result Re-ranking (2009), Learning Designer (2011), Bifrost (2009-2010), Personalized explanations for recommender systems (2006-2009), PIA, News At Hand (2009).
Adaptive Public Displays	Friend Finder, Late-o-Meter, Contextual Display (2010-2011).
Adaptive Museum Visitors Guide	PEACH (2003-2004), APIL (2010).
General Purpose Adaptive Systems	AHA! (1996-2007), GALE (2008-2011), Mouse Smoothing Algorithms for Users with Tremors (2008)

**Table 1-4 Continued**

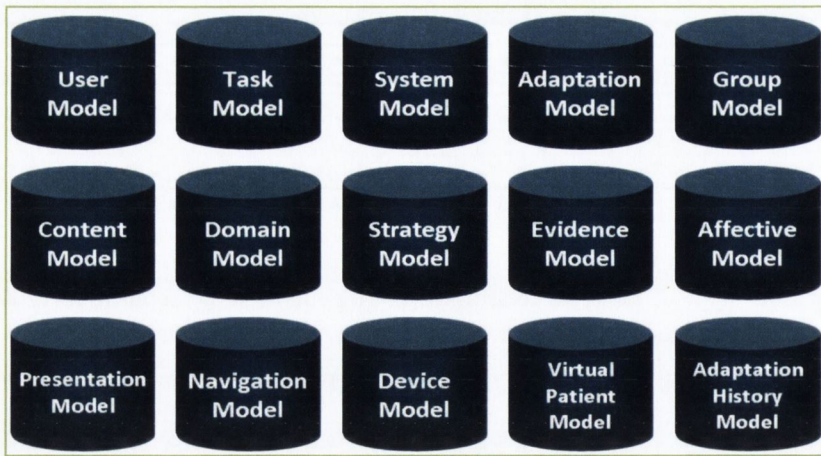
Intelligent Tutoring Systems	Thermo-Tutor (2010)
Adaptive Recommender Systems	Adaptive News System (2005), OWL (2000), ExDis (2011).
Adaptive Educational Game Systems	Crystal Island (2008-2011), ALIGNS (2008-2010).
Adaptive Navigation Systems	Idiotypic control network for a navigating mobile robot (2006-2009)
Adaptive News Systems	Adaptive News System (2005)
Online Help Customer Care Systems	CID (2008), Viper (2008), Adaptive Information Retrieval and Composition System (2010).
Adaptive Training Systems	GeoKaos+Flareqoor (2006-2007).

Given the significance of the tasks performed by AEL systems, in order to produce effective results it is important that evaluation occurs throughout the entire design cycle and provides feedback for design modification (Gena and Weibelzahl, 2007). Furthermore Brusilovsky(20014) argue that, given the large set of techniques and systems, the evaluation and improvement of adaptive systems is more important than inventing new techniques with questionable benefits (Brusilovsky, 2004).

Over the past ten years, developers of adaptive systems have used a variety of internal models during implementation of such systems. A recent real-life user study aimed at investigating existing internal models was performed.<sup>4</sup> The results and findings of this study are discussed in Chapter 7, Section 7.1.3. These results show that most of these systems were developed using the user model; the construction of this model usually requires making many assumptions about the user's skills, knowledge, needs or preferences, as well as about their behaviour and interaction with the system. Figure 2-3 presents a summary of internal models reported by developers and expert evaluators in the study.

---

<sup>4</sup> <http://www.surveymonkey.com/s/Q2DSDF8>



**Figure 2-3: List of internal models used in analysed adaptive systems**

Although the focus of this research is not on evaluation of these internal models, it is important to note that very limited research has been conducted on the evaluations of these models; in particular, there is a lack of reported studies.

The next section presents a summary of AEL systems.

### **2.3.3 Overview of Adaptive E-Learning Systems**

Adaptive E-Learning refers to educational systems that adapt the learning content and the user interface according to the pedagogical and didactical aspects. The aim of these systems is to provide appropriate information to the right student at the right time. Adaptive E-Learning systems (AELS) are able to keep track of usage and to accommodate content automatically for each user and for the best learning result (Esichaikul et al., 2011).

Adaptive E-Learning systems are developed in technology-enhanced learning environments (TELEs) which increasingly offer possibilities for adapting and personalising learning activities and experiences. When technologies are integrated into a single environment or platform to accomplish the goal of enhancing student learning via adaptation, this is referred to as providing an adaptive learning environment. Examples of these environments include adaptive hypermedia, adaptive educational hypermedia, collaborative learning and immersive simulation. These environments provide error feedback that is tailored to the learner or enable the learner to customize the learning environment to fit their interests. Most relevant literature on adaptive learning is focused on adaptivity; by adaptivity is meant the possibility for learners to personalize the course

materials themselves (Burgos et al., 2007). The benefits of adaptivity are seen in E-Learning where users receive personalized guidance through learning material (Conlan et al., 2002, Brusilovsky et al., 2004, De Bra et al., 2003). Such benefits range from increased learning effectiveness to improved user satisfaction (Conlan and Wade, 2004), as well as increased user motivation (Brusilovsky et al., 2004).

Recent research on adaptive E-Learning investigates “how information and communication technologies can be used to support learning and teaching and competence development throughout life” (Svetsky et al., 2010). AEL has attracted much interest with the promise of supporting individual learning tailored to the unique circumstances, preferences and prior knowledge of a learner. Such systems seek to make the E-Learning content more attractive by tailoring it to individual users’ goals and interests. Several researchers have contributed greatly in the AEL field by supporting learning through development of systems that have been acknowledged to have improved and supported the learner’s experience. Examples of such systems are shown in Table 2-4.

Evaluations provide valuable feedback about potential users’ perceptions of the AEL system, how well the software is written, and the extent to which the system really does support decision-making (Jiang and Klein, 1999). In particular, it is important to evaluate the entire AEL system both from a technological perspective and from a user-centred perspective. This is emphasized more in particular in our earlier research on system evaluation (Mulwa et al., 2010). The evaluation of learner and tutor feedback is essential in the production of high-quality personalized TEL services. It is important not only to evaluate the AEL system but also ensure that correct evaluations techniques are used. The reasons for evaluating AEL systems are similar to those for evaluating any type of learning provision, and include determining whether the AEL solution is accomplishing its objectives; identifying who benefits the most or the least from the AEL program, and identifying areas for improvement.

A key challenge faced by evaluators of AEL systems is the difficulty in choosing the right evaluation approach and technique to use. The evaluation of AEL systems is significant due to the complex nature of such systems (Lawless et al., 2010, Tintarev and Masthoff, 2009). Furthermore, many usability issues are raised by end users (Missier Del and Ricci, 2003, Lavie et al., 2005, Weibelzahl and Weber, 2002, Markham et al., 2003) after

interacting with such systems. For example, major challenges include usability issues such as:

- Predictability, transparency, controllability and unobtrusiveness correspond to general usability principles.
- Maintenance of privacy and breadth of experience are relevant to adaptive and personalized E-Learning systems.
- The column of a typical property lists examples of frequently encountered properties of these systems. For example, each has the potential to cause difficulties with respect to one or more of the usability goals.
- The preventive measures aim to ensure that a property is not present in such a manner that it will cause problems, and the compensatory measures goal is to ensure that, in some other way, the goals and objectives are achieved despite the threats created by the properties challenges.

Evaluation of these systems is also a crucial stage in their development (Jameson, 2009). These systems require evaluation due to their inherent usability problems at the interface and to ensure the correctness of adaptive solutions. It is clear that evaluation of AEL systems is difficult and complex (Mulwa C. et.al., 2011, Mulwa C. et. al, 2012, Weibelzahl, 2003). The AEL system reacts differently according to each individual user and the context of use. Evaluation is complex depending on the aspect of personalisation that needs to be evaluated (i.e. 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 (methods, criteria and metrics) need to be combined and executed differently. For example, evaluation of such systems includes: an evaluation of the learner knowledge level at the training session; an evaluation of the learner satisfaction level.

Examples of AEL systems include APeLS (Conlan et al., 2002), AHA! (De Bra and Calvi, 1998), QuizPACK (De Bra, 2002), ELM-ART II (Weber and Brusilovsky, 2001) and JointZone (Ng et al., 2002). These systems build a model of the goals, knowledge and preferences of each individual person and use this model throughout the interaction with the user in order to propose content and link adaptations, which would best suit e-learners. Such systems have been evaluated using a hybrid of evaluation techniques. Table 2-5 presents a summary of AEL systems, the internal models used to develop such systems and

evaluation techniques. The study findings show that reporting of evaluation metrics was very poor compared to reported evaluation methods and criteria

**Table 2- 5: Summary of E-Learning systems evaluation techniques**

Adaptive E-Learning system (2000-2011)		Evaluation techniques		
Name	Internal models	Evaluation methods	Criteria (kind of factors)	Metrics
GATA (2010)	Domain model.	Data mining, Simulated Users.	Knowledge of Domain.	Accuracy of Recommendations.
AdaptWeb (2001)	User model, Domain model, Content model, Presentation model, Navigation model.	Questionnaires, User Observation, Usability Testing, Experimental Evaluation.	*not reported	*not reported
App. Based on AHA! authoring tool	User model, Presentation model, Navigation model.	Interviews, Questionnaires, Data Mining, Usability Testing, User Test.	User Satisfaction, Content Adaptation, User Performance.	UiAI: User Interaction Adaptivity Index, pIA: Performance Influence on Adaptivity, ApOC: Adaptive Personalisation Overall Cost.
Peer Finder	User model, Group model	Eye-Tracking, Task Completion Time, System Preference Survey.	*not reported	*not reported
NetCoach (2000)	User model, content model, presentation model, navigation model	Questionnaires	Usability, Perceived Usefulness, Appropriateness of Adaptation, Usability of Interface Adaptation, User Satisfaction.	*not reported

Table 2-5 Continued

Adaptive E-Learning system (2000-2011)		Evaluation techniques		
Name	Internal models	Evaluation methods	Criteria (kind of factors)	Metrics
OWL (2000)	User model, Domain model.	Interviews, Questionnaires, Prototyping, Expert Review, Usability Testing, Experimental Evaluation, Empirical Observations, Quantitative.	Usability, Perceived Usefulness, Trust and Privacy Issues, User Behaviour, User Satisfaction, Piloting.	*not reported
aLFanet (2002- 2005),	User model, content model, educational standards (IMS family)	Interviews, Questionnaires, Focus Group, User Observation, Data Mining, Prototyping, Wizard of Oz Simulation, Scenario Based Design, Usability Testing, Empirical Observations, Quantitative.	Usability, Perceived Usefulness, Intention to Use, Appropriateness of Adaptation, User Behaviour, User Goal, User Satisfaction, Early Prototype Evaluations, Preferences, User Performance.	*not reported

The next section discusses adaptive hypermedia systems, which are the most common variation type (category) of AEL systems.

### 2.3.3.1 Adaptive Hypermedia Systems

Over the past fifteen years, adaptive hypermedia systems (AHS) have been increasingly employed for educational purposes, especially with the emergence of distance and distributed learning (Davies, 1999). A core element of an AH system is the user model. Several researchers have used AH techniques to provide a personalized learning

experience that draws on computer-driven intelligent tutoring systems (ITSs) and student-driven virtual learning environments (VLEs), which are dedicated learning environments. A fundamental tenet of education is that students are different and hence learn in different ways (Brown, 2007, Jones, 1996). Brown defines a VLE as a software system that helps to manage a computer-based learning course; sometimes the term ‘managed learning environment’ (MLE) is used. Most VLEs develop a number of elements such as course syllabus, administrative information, student registration, tracking facilities and teaching materials (such as course content) and additional learning resources (such as reading lists and links to resources on the Internet). Such environments are known to facilitate and have the ability to provide multi-choice quizzes, which are scored automatically; communication tools (i.e. emails, bulletin boards, chat rooms), and also to produce course statistics and documentation on the usage and performance of the system. However, VLE environments also have problems. Brown (2007) identified a number of issues; for example, none of the commercial or open-source VLEs (i.e. Moodle,<sup>5</sup> BlackBoard,<sup>6</sup> WebCT<sup>7</sup>) provided any kind of adaptation to support the variety in characteristics (such as cognitive preferences or motivation) shown by different users. Furthermore the researcher suggests that, in order to provide adaptivity to users, the system’s source code needs to be modified; it was either not available to developers or was integrated into other parts of the software and hence difficult to debug.

Currently the hypermedia system or application offers learners much freedom to navigate through a large hyperspace. On the other hand, adaptive hypermedia (AH) offer learners personalized content, presentation and navigation support. Knutove et al. (2009) provide a comprehensive overview of AH methods and techniques since their introduction 12 years ago (Knutov et al., 2009). In this research an adaptive hypermedia system (AHS) is defined as “any hypertext and hypermedia system which reflects some features of the user in a user model and applies this model to adapt various visible aspects of the system to the user (Brusilovsky, 1996)”. In other words, an AH system should be able to satisfy three criteria: it should be a hypertext or hypermedia system; it should have a user model; and it should be able to adapt the hypermedia using this model. Most AH systems exceed these basic criteria by adding multiple models.

---

<sup>5</sup> <http://moodle.org/>

<sup>6</sup> <http://www.blackboard.com/>

<sup>7</sup> <http://www.webct.com/>



O’Keeffe et al. (2006) note that, in the past, AHS systems attempted to customize courses to a learner’s prior knowledge, goals and personal preferences without taking into account any form of pedagogy. As a result, such systems neglected the entire body of research that exists in the educational field and failed to take advantage of the benefits that the application of pedagogy has for the learning experience (O’Keeffe et al., 2006). Furthermore, in a review Karampiperis et al. (2005) identified the current state-of-the-art adaptive hypermedia systems as AHA! (De Bra et al., 2002), OntoAIMS (Aroyo et al., 2003), the Personal Reader (Dolog et al., 2004), WINDS (Kravcik and Specht, 2004), ACCT (Dagger et al., 2005), which are based on the adaptive hypermedia application model (AHAM). This model builds upon the Dexter model, a common model for hypertext-based systems that was designed for general-purpose adaptive web application. The model consists of two main layers: the run-time layer, which contains the adaptation engine that performs the actual adaptation, and the storage layer, which stores information about the media space, the domain model, the user model and the adaptation model (Karampiperis and Sampson, 2005).

The most common hypermedia systems are adaptive educational hypermedia systems (AEHSs). The next section presents an overview of these.

### ***2.3.3.2 Adaptive Educational Hypermedia Systems***

Educational hypermedia was one of the first application areas of adaptive hypermedia and is currently one of the most popular and well-researched and investigated areas of research. Adaptive educational hypermedia systems (AEHSs) were developed to address learner dissatisfaction through personalizing the learning experience. The typical architecture of the state-of-the-art AEHS system, which is fully decoupled, consists of five complementary models: the domain model, which specifies what is to be adapted; the user and context models, which indicate the parameters for the adaptation of content, and the instructional and adaptation models, which express the pedagogical approach the learning process should be based on, as well as the forms of adaptation to be performed (Karampiperis and Sampson, 2005). Figure 2-4 depicts the generalized architecture of an AEH system.

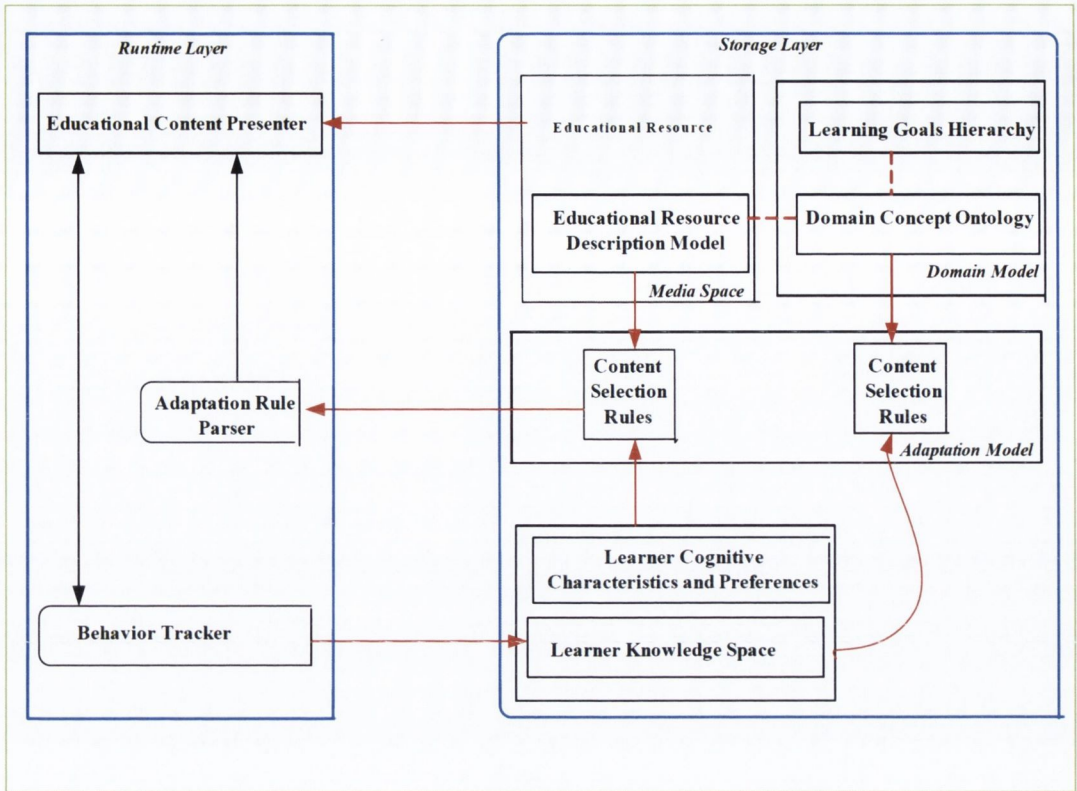


Figure 2-4: Generalized architecture of an adaptive educational system

The generic layers (knowledge representation, adaptation and interface) in a simplified example of the adaptive educational hypermedia architecture are depicted in Figure 2-5.

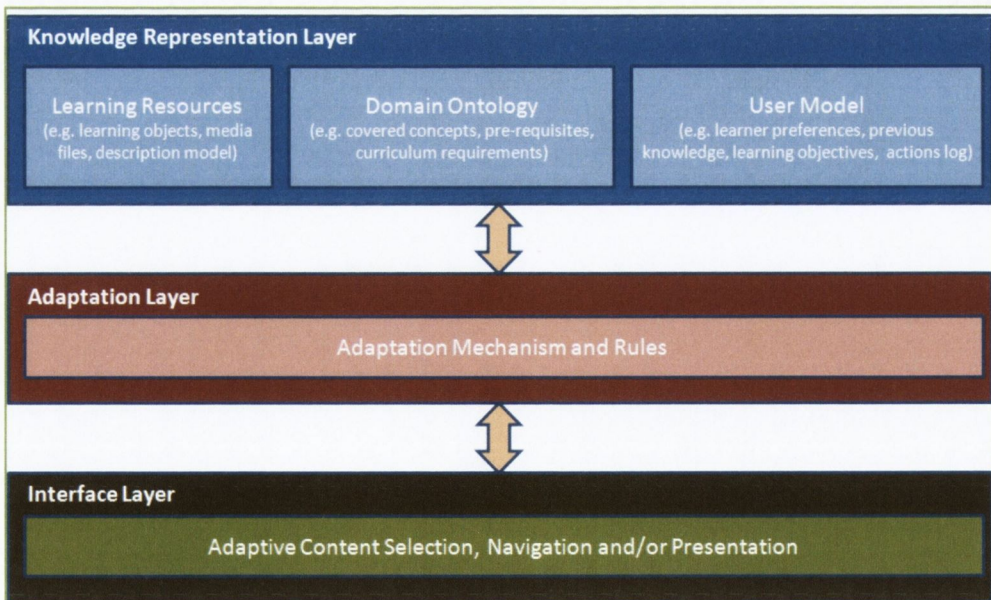
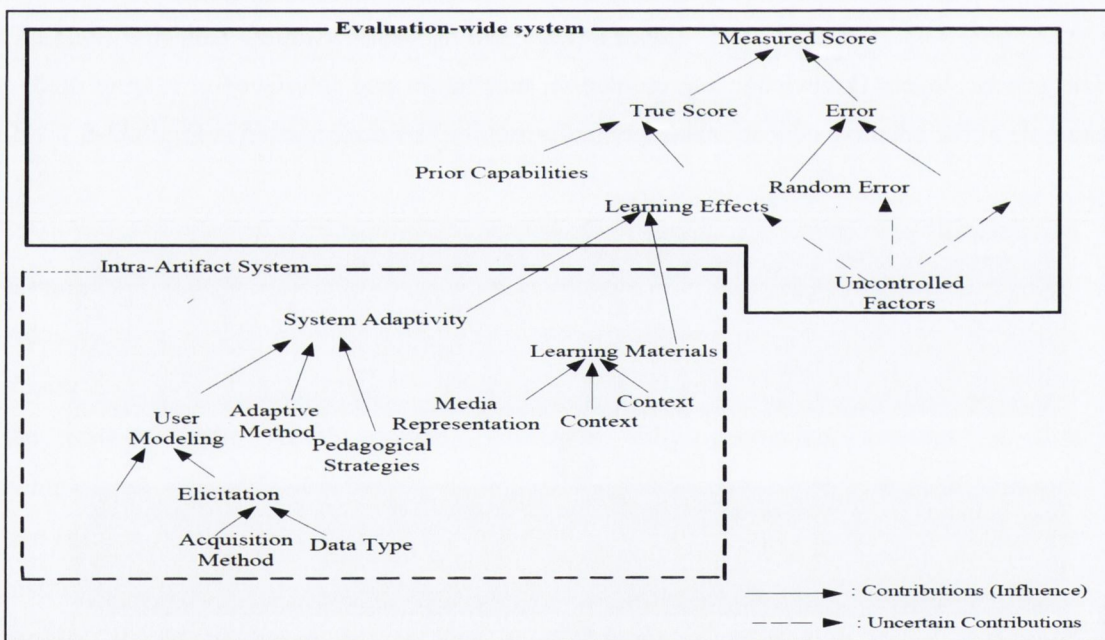


Figure 2-5: Generic layers in simplified example architecture of an educational AEH

The purpose of such adaptive educational offerings is to maximize learner satisfaction, learning speed (efficiency) and educational effectiveness (Popescu et al., 2007). AEHSs have been found to be useful in engaging the learner more in the educational experience. AEHS systems offer an alternative to the non-individualized instruction approach by providing various services adapted to the learner profile. These systems are based on user models that characterise each individual and can use these models to offer learners educational experiences that fit their needs. To achieve this, AEHSs are comprised of several sub-components which have their own distinct behaviours and properties. Figure 2-6<sup>8</sup> conceptually depicts the hierarchy of this scenario. For example, the AEHS as the intra-artefact system shown in Figure 2-6 can be decomposed by considering the influence of sub-components on the performance of ones that are at a higher level. The dashed arrows are edges representing uncertain influence. In this model the uncontrolled factors are identified and linked to other sub-components in the system with dashed arrows. For example, sub-components inside the intra-artefact system can play the role of uncontrolled factors in the evaluation. On the other hand, the hierarchy of the evaluation task is shown by the evaluation-wide system.



**Figure 2-6: Hierarchy of underlying factors of AEH (Mulwa et al., 2010)**

Brusilovsky (2004) provided a review of past and present research on AEH systems. The researcher categorized AEH systems into three generations, which can be traced back to

<sup>8</sup> This diagram is taken from <http://nccur.lib.nccu.edu.tw/bitstream/140.119/14993/1/52.pdf>

the early 1990s, while the second generation evolved from 1996 on. Jovanovica et al. (2009) discuss the three generations of AEHSs. Both researchers acknowledge that the first generation of AEH systems comprised stand-alone systems with adaptation rules and content entwined in a single model. They used this model together with the user model to offer personalized content (AHA! and ELM-ART). However, as adaptation rules and content were intertwined, there was little scope for content or model reuse or the use of externally developed content in the generation of learning offerings.

The second generation attempted to overcome some of the problems encountered by the first generation by pursuing a multi-model approach (Jovanović, 2009). This approach assumed decoupling of content and the adaptation rules of the system (Conlan, 2005). The third generation currently is moving towards a service-oriented architecture and the complete decoupling of different kinds of knowledge (Jovanovica et al., 2009). Brusilovsky (2004) provided a subjective overview of research in adaptive educational hypermedia and summarized the current state of the art of the three generations (Brusilovsky, 2004). The researcher acknowledges that problems were encountered while using the AEHS and accepted that several research teams had recognized the problems of static hypertext in different application areas, and had begun to explore various ways to adapt the behaviour of hypertext and hypermedia systems to users. For example, he accepts there are problems related to hypermedia such as navigation in hypermedia, inefficient navigation or the problem of being lost in hyperspace; such problems were discovered when the field of hypertext reached relative maturity at the end of the 1980s (Brusilovsky, 2004).

Recent research shows that there are few evaluations available in the AH domain relative to the amount of research interest this domain is attracting. Most of the research in this domain focuses on the technological design and performance of systems without justifying the designs through the lessons learned from evaluations (Conlan and Wade, 2004). To provide the best support for learners, a user-centred evaluation approach for enhancing and validating the student model of AEHS has been proposed, combining adaptive hypermedia (AH) and information retrieval techniques (Séamus Lawless et al., 2010).

A real-life user study was conducted aimed at investigating how AEH systems developed from 2000 to 2011 had been evaluated. The results of this study are depicted in Table 2-6. The findings of this study show that there are tradeoffs (cross-over) between evaluation

techniques (methods, criteria and metrics), especially with evaluation techniques used in user-centred evaluation approaches and layered evaluation approaches.

**Table 2- 6: Summary of evaluations of adaptive educational hypermedia systems (2000 to 2012)**

Adaptive Educational Hypermedia		Evaluation Technique(s)		
System Name	Internal models	Methods	Criteria	Metrics
UNITE (2008-2011)	User Model, Domain Model, Content.	Interviews, Questionnaires.	Usability, Perceived Usefulness, Intention to Use, Trust and Privacy Issues, Appropriateness of Adaptation, User Behaviour, User Goal, Knowledge of Domain, User Satisfaction, Interface Knowledge, Content Adaptation, User Performance, Transparency, User Cognitive Workload.	Accuracy of Recommendations.
SIETTE (1998-2011)	User Model, Domain Model	Questionnaires, Data Mining, Simulated Users, Cross-Validation, Usability Testing, Experimental Evaluation, User Test, Creative Brainstorming Sessions, Empirical Observations, Empirical Observations, Quantitative, Grounded Theory.	Usability, Perceived Usefulness, Intention to Use, User Behaviour, Usability of Interface Adaptation, User Satisfaction, User Performance.	Accuracy of Recommendations, Reliability Metrics, Precision.

Table 2-6 Continued

Adaptive Educational Hypermedia		Evaluation Technique(s)		
System Name	Internal models	Methods	Criteria	Metrics
MyPlan(2008) Migen (2010) Inspire (2000)	User Model, Content Model, Presentation Model, Task Model, Strategy Model, Navigation Model	Interviews, Questionnaires, User Observation, Simulated Users, Heuristic Evaluations, Prototyping, Expert Review, Wizard of Oz Simulation, Scenario Based Design, Usability Testing, User Test	Usability, Perceived Usefulness, Intention to Use, Appropriateness of Adaptation, User Behaviour, User Satisfaction, Early Prototype Evaluations, Content Adaptation, Real User Actions, to Combine Qualitative Evaluation.	Accuracy of Recommendations, Accuracy of Retrieval, Reliability Metrics, Precision, pQoR: Performance Quality on Response.
Dashboard at KiWi Framework (2007)	User Model, Navigation Model	User Observation, Usability Testing, User Test.	Usability, User Behaviour, User Goal, Interface Knowledge, Content Adaptation	Accuracy of Recommendation, Accuracy of Retrieval, Precision
MEDEA (2006)	User Model, Domain Model	Heuristic Evaluations, Prototyping, Expert Review, Experimental Evaluation, Empirical Observations, Formative, Summative evaluation.	Contents Reutilization Capabilities, Syntactic and Semantic Interoperability.	*not reported

Table 2-6 Continued

Adaptive Educational Hypermedia		Evaluation Techniques		
System Name	Internal models	Methods	Criteria	Metrics
CoMoLE (2007-2008)	User Model, Domain Model, Content Model, Presentation Model, Device Model, Task Model, Strategy Model, Navigation Model.	Interviews, Questionnaires, User Observation, Data Mining, Expert Review.	Usability, Perceived Usefulness, Intention to Use, Appropriateness of Adaptation, User Behaviour, User Goal, Appropriateness of Adaptation, User Behaviour, User Goal, Knowledge of Domain, Background and Hyperspace Experience, Usability of Interface Adaptation, User Satisfaction, Early Prototype Evaluations, Evaluation before Implementation, Content Adaptation, Preferences, User Skills and Capabilities, User Performance, Real User Actions, To Combine Qualitative Evaluation, Collaboration with Real Users During Final Evaluation Step.	Accuracy of Recommendations, Accuracy of Retrieval, UiAI: User Interaction Adaptivity Index, pQoR: Performance Quality on Response, pIA: Performance Influence on Adaptivity, AvgpACF: Average Personalisation Adaptive Cost Per Functionality, ApOC: Adaptive Personalisation Overall Cost, DSAI: Domain Specific Adaptivity Index.

Table 2-6: Continued

Adaptive Educational Hypermedia		Evaluation Techniques		
System Name	Internal models	Methods	Criteria	Metrics
iOLMlets (2006)	User Model	Questionnaires, User Observation, Data Mining, Expert Review, Experimental Evaluation, User Test, Creative Brainstorming Sessions, Empirical Observations, Quantitative.	Perceived Usefulness, Intention to Use, Trust and Privacy Issues, Appropriateness of Adaptation, User Behaviour, User Goal, Knowledge of Domain, Early Prototype Evaluations, Preferences, User Skills and Capabilities, User Performance, Transparency, Appropriateness, Real User Actions, To Combine Qualitative Evaluation.	*not reported
ActiveMath (2000-2011)	User Model, Domain Model, Content Model, Presentation Model, System Model, Strategy Model, Navigation Model.	Interviews, Questionnaires, User Observation, Data Mining, Simulated Users, Cross-Validation, Wizard of Oz Simulation, Usability Testing, Experimental Evaluation, User Test, Empirical Observations, User Test, Empirical Observations, Quantitative, Grounded Theory.	Usability, Perceived Usefulness, Intention to Use, Appropriateness of Adaptation, User Satisfaction, Content Adaptation, User Performance, Real User Actions.	Precision



**Table 2-6: Continued**

Adaptive Educational Hypermedia		Evaluation Techniques		
System Name	Internal models	Methods	Criteria	Metrics
www.assistment.org (2002)	User Model, Domain Model, Strategy Model, Navigation Model.	Interviews, Questionnaires, Focus Group, Data Mining Simulated Users, Cross-Validation, Heuristic. Evaluations, Expert Review, Usability Testing, Experimental Evaluation, Empirical Observations, Cooperative Evaluation	Usability, Perceived Usefulness, Intention to Use, Knowledge of Domain, Background and Hyperspace Experience, Usability of Interface Adaptation, Interface Knowledge, User Performance, Real User Actions, Collaboration with Real Users during Final Evaluation Step.	*not reported
SHAIEX (2002)	User Model, Content Model, Task Model.	Interviews, Questionnaires, User Observation, Prototyping.	Usability, Intention to Use, Appropriateness of Adaptation, User Behaviour, User Satisfaction, Early Prototype Evaluations, Evaluation before Implementation, Real User Actions, Collaboration with Real Users during Final Evaluation Step.	Accuracy of Recommendations
Adaptive extension to an E-Learning platform (2008-2010)	User Model, Domain Model, Content Model, Task Model.	Interviews, Questionnaires, Focus Group, Data Mining, Usability Testing.	Usability, Trust and Privacy Issues, User Behaviour, User Goal, Usability of Interface Adaptation, Content Adaptation, User Performance.	Accuracy of Retrieval, Behavioural Complexity, Precision.

Several researchers have identified major challenges encountered by evaluators of AEL systems. Manouselis et al. (2011) identified two major difficulties. First, adequately defining the reference variables against which the adaptivity of the system will be evaluated is difficult 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 (Manouselis et al., 2011). Secondly, criteria for the success of adaptivity are not well defined or there are rarely commonly accepted criteria. On the one hand, objective standard criteria (e.g. duration, number of interaction steps, knowledge gain) regularly failed to find a difference between adaptive and non-adaptive versions of a system. On the other hand, subjective criteria that are standard in human-computer interaction research (e.g. usability questionnaires) have been rarely applied to measure the success of adaptive systems. In TEL the issues are related to the definition of appropriate evaluation methods (e.g. techniques, metrics and instruments) to measure the success of a successful recommendation strategy in comparison to a non-successful one (Manouselis et al., 2011).

To tackle the identified challenges and issues, this research proposes an evaluation framework for supporting novice and expert evaluators of adaptive systems (EFEx). Chapter 3 and 4 present a detailed description of this framework.

### ***2.3.3.3 Intelligent Tutoring Systems***

Research in the field of intelligent tutoring systems (ITSs) indicates that they have proven their effectiveness not only in controlled lab studies but also in real classrooms. These ITSs are computer-based instructional systems with models of instructional content that specify *what* to teach, and teaching strategies that specify *how* to teach (Murray, 2003). Such systems make inferences about a student's mastery of topics or tasks in order to dynamically adapt the content or style of instruction.

Several researchers acknowledge that ITSs have proven their effectiveness in classrooms (Koedinger et al., 1997, Mitrovic and Ohlsson, 1999, Mitrovic et al., 2004, Mitrovic et al., 2007, Mitrovic et al., 2008). These systems achieve important improvements in comparison to classroom learning, due to the knowledge about the instructional domain, pedagogical strategies and student modelling capabilities.

A key challenge faced by users of these systems is that they still have not achieved widespread effect on education due to their high complexity and difficulty of development. Murray (2003) indicates that composing the domain knowledge required for ITSs consumes most of the total development time (Murray, 2003). The researcher emphasizes that this task requires multi-faceted expertise, including knowledge engineering, artificial

intelligence (AI) programming and the domain itself. The researcher further emphasizes the need for more empirical testing using multiple authors and domains; more research on authoring student models; more complete and standardized ontologies and metadata standards; more research on the differential effectiveness of various computationally explicit instructional strategies; and more exploration of open component-based architectures.

The findings of a real-life user study conducted on evaluations of adaptive ITS systems developed from 2000 to 2012, show that there is tradeoff between the different evaluation approaches used and evaluation techniques. Table 2-7 presents a summary of these findings (system name, internal models used when developing systems and evaluation techniques).

**Table 2- 7: Summary of intelligent tutoring systems, internal models and evaluation techniques**

Adaptive Tutoring Systems		Evaluation Techniques		
System Name	Internal models	Methods	Criteria	Metrics
SQL-Tutor (1998-2011), EER-Tutor (2003), NORMIT (2002), UML-Tutor (2004), ERM-Tutor (2005), J-Latte (2006), CID (2008), Viper (2008)	User Model, Domain Model, Task Model, Affective Model.	User Observation, Data Mining, Simulated Users, Wizard of Oz Simulation, Usability Testing, Experimental Evaluation.	Usability, Perceived Usefulness, Intention to Use, Appropriateness of Adaptation, Usability of Interface Adaptation, Early Prototype Evaluations, To Combine Qualitative Evaluation.	*not reported

Table 2-7 Continued

Adaptive Tutoring Systems		Evaluation Techniques		
System Name	Internal models	Methods	Criteria	Metrics
EMSAVE (2009), MOPET Track-Rate (2008-2011), Monster & Gold (2008-2010), MOPET (2006-2008), Geokaos (2006-2007)	User Model, Content Model, Task Model.	Interviews, Questionnaires, Focus Group, User Observation, Expert Review, Usability Testing, Experimental Evaluation, User Test, Empirical Observations, Quantitative.	Usability, Perceived Usefulness, Intention to Use, Knowledge of Domain, User Satisfaction, Preferences, User Skills and Capabilities, User Performance, Real User Actions, Collaboration with Real Users during Final Evaluation Step.	*not reported
JTS: A Multi-Agent Java tutoring system (2001), ViSMod: Interacting with Bayesian student models (2002), The Learning Game (2003) BELLA/English-Math ABLE (EM-ABLE) (2007-2008)	User Model, Domain Model, Content Model, Presentation Model, Task Model, Strategy Model, Navigation Model, Evidence Model.	User Observation, Verbal Protocol, Expert Review, Usability Testing, Experimental Evaluation, Empirical Observations, Quantitative.	Usability, Trust and Privacy Issues, Appropriateness of Adaptation, User Satisfaction, User Skills and Capabilities, User Performance, User Cognitive Workload.	Accuracy of Recommendations, pIA: Performance Influence on Adaptivity, ApOC: Adaptive Personalisation Overall Cost.

#### *2.3.3.4 Adaptive Educational Game Systems*

Recent research in the field of adaptive educational game (AEG) systems shows that they have focused on the learning benefits provided by the inherent motivation, rich visualizations and low risk of failure provided by contemporary educational games (Peirce and Wade, 2010). Although these systems create highly engaging and immersive learning environments, additional techniques can further aid the learning process. The integration of personalisation into educational games presents challenges beyond those faced in ITS or AH systems. The main challenge is that the objectives of instructional design and engaging gameplay can conflict. This evidently requires compromises in either gameplay or learning personalisation. As identified by a number of authors, an educational game must be a game first and learning tool second (Van Eck, 2007; Prensky, 2001); similarly an adaptive game must be a game first and a personalized learning experience second.

Despite their potential benefits, the instances of adaptive educational games are scarce. One possible cause is the complexity of integrating personalized learning into a gaming environment. Although the research area of adaptive educational games is still emerging, strides are already being taken in the variety, complexity and reusability of the adaptation provided. With EU projects such as ELEKTRA and 80Days (80Days-Project n.d.), there is evidently a growing interest in adaptive educational games. Kickmeier-Rust et al. (2007) states that the intrinsic motivation to play, and therefore to learn, that might be provided by digital educational games “teases researchers and developers”, adding: However, existing educational games often fail in their attempt to compete with commercial games and to provide successful learning. Often some learning is added to digital games or some gameplay is added to educational applications. Successful educational games, however, require merging professional game design with sound pedagogical strategies, creating a new hybrid format” (Kickmeier-Rust et al., 2007).

Immersive educational computer games offer a highly promising approach to overcoming the weaknesses mentioned above and to make learning more engaging, satisfying, and probably more effective. Currently, there is much hype about game-based learning, ranging from entertainment games to games with primarily educational purposes. An overview of these games is presented by (Mitchell and Savill-Smith, 2004). They state: “The major strength of digital games in education is a high level of intrinsic motivation to play and to

proceed in the game and, thus, to learn within the context of a meaningful and continuous storyline and within the related parasocial dimension provided by game characters.”

According to Malone (Malone, 1981), the factors that provide that strength include the games being fun, involving fantasy, and arousing curiosity. Educational games provide clear goals and rules, a meaningful learning context, an engaging storyline, immediate feedback, a high level of interactivity, challenge and competition, random elements of surprise, and rich and appealing learning environments (Malone, 1981). These factors determine the motivation to play and to learn but are also considered to be important for successful and effective learning (interactivity, feedback, problem-solving, or context effects). (Merrill, 2002, Schulmeister, 2004) present a review of these factors. Major disadvantages of these systems include difficulties in providing an appropriate balance between gaming and learning activities, providing a continuous balance between challenge and ability, aligning the game with national curricula, and the extensive costs of developing high-quality games (Van Eck, 2006). Van Eck concludes that, due to these problems, most of today’s educational games cannot compete with their commercial counterparts in terms of gaming experience, immersive and interactive environments and storytelling, or intrinsic motivation to play.

Pierce et al. (2008) emphasize that educational games have the potential to provide intrinsically motivating learning experiences that immerse and engage the learner. However, “the much-heralded benefits of educational games” seldom take into consideration “the one-size-fits-all approach to education they typically embody” (Pierce et al., 2008). “The potential presented by adaptive educational games heralds an era of motivating personalized learning experiences. Such an advance nevertheless must overcome the conflicts exposed by adapting a gaming experience for educational gain”.

Educational games can be seen as a progression in technology-enhanced learning that provides direct support for a learner’s motivation (Rieber, 1996). Although games can provide an intrinsically motivating experience, the complexities of educational game design are considerable (Akilli, 2007). With the full potential of educational games yet to be realized (Van Eck, 2007), one must consider the existing approaches to technology-enhanced learning that have proven fruitful. For instance, “the stalwart of adaptation has long proven beneficial in eLearning as is evident in Adaptive Hypermedia” (Brusilovsky, 1996). Combining adaptation and educational games can uniquely present a personalized

supportive motivational experience. In realizing this motivation through appropriate challenge, curiosity, fantasy and control (Conlan and Wade, 2004) there remains great potential to address the under-motivated learner.

Without an immersive gaming experience, the benefit of using games as a motivational vehicle for learning becomes compromised. A number of authors (Van Eck, 2007, Papert S., 1998) have identified that an educational game must be a game first and an educational tool second. Without this prioritization the potential benefits of gaming are reduced. Although there has been steady growth in the variety and complexity of educational games available, the instances and quality of adaptive educational games remain limited. Integrating personalized learning experiences into educational games raises significant challenges for an area of research that is only now making progress away from earlier 'Shavian Reversals' (Papert S., 1998). In many instances little was done to blur the boundaries between gaming and learning, something which is considered a desirable feature<sup>9</sup>. Whilst research into the effective integration of gaming and learning is ongoing, the compulsion to provide a personalized educational experience is driven by established research in Intelligent Tutoring Systems (ITSs) (Pivec, 2007) and Adaptive Hypermedia (Van Eck, 2006). The provision of personalisation is long known to be beneficial to learning outcomes and experience (Brusilovsky, 1996). The potential thus remains that, through personalisation of the learning experience in educational games, the intrinsic motivation provided by games can be complemented with a tailored learning experience.

Such games as the DARPA-funded Tactical Language and Cultural Training System (TLCTS) (Brusilovsky, 2001) have shown that effective learning outcomes can be achieved through the use of adaptive educational games (Brusilovsky, 2001). One of the key motivational factors found in games is a strong storyline (Johnson et al., 2007), a tool which is often used in educational games. While integrating educational content within a motivating narrative is a challenging task, the further complexity added when considering adaptive educational content is one of the considerable challenges facing adaptive educational games. The instances of adaptive educational games continue to increase but they are still few in comparison to the growing number of non-adaptive educational games.

Conlan et al. (2009) emphasize that digital educational games (DEGs); offer immersive environments through which learners can enjoy motivational and compelling educational

---

<sup>9</sup> ELEKTRA - Enhanced Learning Experience and Knowledge TRAnsfer", Retrieved from <http://www.elektra-project.org> on 25th June 2008

experiences. Applying personalization techniques in these games can further enhance the educational potential, but the often real-time and narrative-driven focus of games presents many challenges to traditional adaptation approaches (Conlan et al., 2009). DEGs are acknowledged as an emerging area in which personalization techniques, traditionally developed within the Adaptive Hypermedia (AH) research domain, are being applied. A major issue that has plagued online learning solutions for quite some time has been the high levels of drop-out (Frankola, 2001), often precipitated by poor intrinsic motivation and relevance in the material presented. To offer appropriate adaptive interventions three challenges must be overcome:

- Modelling of the learner's knowledge acquisition (also referred to as cognitive gain) must be achieved in real time.
- Adaptive hypermedia techniques, which are typically applied to web-based systems, also need to operate in real time.
- The personalizations offered must not adversely affect the flow (Csikszentmihalyi M., 1990) of the game. The challenges of real-time adaptation and the maintenance of flow (Csikszentmihalyi M., 1990) stem from the need to maintain a learner's immersion in the gaming experience.

Adaptive Hypermedia Systems have typically dealt with narrative from a different perspective. The most prevalent examples come from the adaptive E-Learning domain where narrative usually refers to the flow of a piece of coursework (Specht, 2002, De Bra et al., 2003). A summary of evaluations of AEGs systems is presented in Table 2-8.

**Table 2- 8: Summary of evaluations of adaptive educational games**

Adaptive Educational Games		Evaluation Techniques		
Systems Name	Internal models	Evaluation Methods	Criteria	Metrics
Crystal Island (2008-2011)	User Model, Domain Model, Content Model, Task Model, Strategy	Questionnaires, Focus Group, User Observation, Data Mining, Cross-Validation, Prototyping, Expert Review, Wizard of Oz Simulation,	Usability, Perceived Usefulness, Appropriateness of Adaptation, User Behaviour, User Goal, Knowledge of Domain, Background and Hyperspace Experience, User	Precision, Recall, and Accuracy of models, Student responses to the software, including impacts on students' Content learning gains, Presence Questionnaire Scores, Intrinsic motivation inventory



	Model.	Usability Testing, Experimental Evaluation, User Test, Creative Brainstorming Sessions, Empirical Observations.	Satisfaction, Interface Knowledge, User Skills and Capabilities, User Performance, Real User Actions.	scores, Gameplay characteristics, Problem-solving performance, and Self-reported moods.
ALIGN (2008-2010)	User Model, Adaptation History Model, Game State Model.	Interviews, Questionnaires, Discussion Group, User Observation, Data Mining, Usability Testing, Experimental Evaluation, User Test, Creative Brainstorming Sessions, Empirical Observations, Quantitative.	Usability, Perceived Usefulness, Appropriateness of Adaptation, User Behaviour, User Satisfaction, Content Adaptation, Preferences, User Skills and Capabilities, User Performance.	Perceived appropriateness of adaptations, Invasiveness of adaptations, Awareness of adaptations.

### 2.3.3.5 Adaptive Educational Hybrid Recommender Systems

Recommender systems were originally defined as ones in which “people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients” (Hernández del Olmo and Gaudioso, 2008). Now, a broader and more general definition is being adopted in the field, referring to recommender systems as those systems that “have the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options” (Burke, 2002).

Hybrid recommender systems combine two or more recommendation techniques (collaborative, content-based, utility-based, knowledge-based and case-based) in order to gain improved performance, with fewer of the drawbacks of any individual one. For example, most commonly collaborative filtering combines with other techniques in an

attempt to avoid ramp-up problem<sup>10</sup>. Table 2-9 presents existing tradeoffs between recommendation (combinations) techniques which have been used by developers of hybrid recommender systems and examples of current strengths and weakness.

**Table 2- 9: Tradeoffs between recommendation techniques (Burke, 2002)**

Technique	Background	Input	Process
Collaborative	Ratings from $U$ of items in $I$ .	Ratings from $u$ of items in $I$ .	Identify users in $U$ similar to $u$ , and extrapolate from their ratings of $i$ .
Content-based	Features of items in $I$	$u$ 's ratings of items in $I$	Generate a classifier that fits $u$ 's rating behavior and use it on $i$ .
Demographic	Demographic information about $U$ and their ratings of items in $I$ .	Demographic information about $u$ .	Identify users that are demographically similar to $u$ , and extrapolate from their ratings of $i$ .
Utility-based	Features of items in $I$ .	A utility function over items in $I$ that describes $u$ 's preferences.	Apply the function to the items and determine $i$ 's rank.
Knowledge-based	Features of items in $I$ . Knowledge of how these items meet a user's needs.	A description of $u$ 's needs or interests.	Infer a match between $i$ and $u$ 's need.

Several developers of hybrid recommender systems have attempted to combine several methods when developing such systems. Table 2-10 presents some of the combination methods that have been employed. Following is a brief description of these methods:

- *Weighted*: A weighted hybrid recommender is one in which the score of a recommended item is computed from the results of all the available recommendation techniques present in the system. A potential benefit of a weighted hybrid is that all of the system's capabilities are brought to bear on the recommendation process in a straightforward way and it is easy to perform post-hoc credit assignment and adjust the hybrid accordingly.

<sup>10</sup> This term describes an issue with the recommendation systems (new items cannot be recommended to any user until they get some sort of rating). Recommendations for items that are new to the database are essentially relatively weaker than more widely rated products, and this is the same case for users who are new to the system.

**Table 2- 10: Hybridization (combination) of recommendation methods  
(Burke, 2002)**

Hybridization method	Description
Weighted	The scores (or votes) of several recommendation techniques are combined together to produce a single recommendation.
Switching	The system switches between recommendation techniques depending on the current situation.
Mixed	Recommendations from several different recommenders are presented at the same time
Feature combination	Features from different recommendation data sources are thrown together into a single recommendation algorithm.
Cascade	One recommender refines the recommendations given by another.
Feature augmentation	Output from one technique is used as an input feature to another.
Meta-level	The model learned by one recommender is used as input to another.

- *Switching*: A switching hybrid builds in item-level sensitivity to the hybridization strategy: the system uses some criterion to switch between recommendation techniques. For example, the DailyLearner system uses a hybrid of content and collaborative in which a content-based recommendation method is employed first. If the content-based system cannot make a recommendation with sufficient confidence, then a collaborative recommendation is attempted.<sup>11</sup> This switching hybrid does not completely avoid the ramp-up problem, since both the collaborative and the content-based systems have the ‘new user’ problem. What the collaborative technique provides in a switching hybrid is the ability to cross genres, to come up with recommendations that are not close in a semantic way to the items previous rated highly, but is still relevant.
- *Mixed*: Where it is practicable to make a large number of recommendations simultaneously, it may be possible to use a ‘mixed’ hybrid, where recommendations from more than one technique are presented together. The PTV system (Smyth and Cotter, 2000) uses this approach to assemble a recommended programme of television viewing. It uses content-based techniques based on textual descriptions of TV shows and collaborative information about the preferences of other users. Recommendations from the two techniques are combined in the final suggested programme. The mixed hybrid avoids the ‘new item’ start-up problem; the content-based component can be relied on to recommend new shows on the basis of their descriptions even if they have not been rated by anyone.

<sup>11</sup> Actually, the Billsus system has two content-based recommendation algorithms, one short-term and one long-term, and the fallback strategy is short-term/collaborative/long-term.

- *Feature Combination*: Another way to achieve the content/collaborative merger is to treat collaborative information as simply additional feature data associated with each example and use content-based techniques over this augmented dataset. The feature combination hybrid lets the system consider collaborative data without relying on it exclusively, so it reduces the sensitivity of the system to the number of users who have rated an item. Conversely, it lets the system have information about the inherent similarity of items that are otherwise opaque to a collaborative system.
- *Cascade*: One technique is employed to produce a rating or classification of an item and that information is then incorporated into the processing of the next recommendation technique. For example, the Libra system (Mooney and Roy 1999) makes content-based recommendations of books based on data found in Amazon.com, using a naïve Bayes text classifier. The text data used by the system includes ‘related authors’ and ‘related titles’ information that Amazon generates using its internal collaborative systems. These features were found to make a significant contribution to the quality of recommendations.
- Augmentation is attractive because it offers a way to improve the performance of a core system, like Net Perceptions’ GroupLens Recommendation Engine or a naïve Bayes text classifier, without modifying it.
- Another way that two recommendation techniques can be combined is by using the model generated by one as the input for another. This differs from feature augmentation: in an augmentation hybrid a learned model is used to generate features for input to a second algorithm; in a meta-level hybrid, the entire model becomes the input. The first meta-level hybrid was the web filtering system Fab (Balabanovic 1997, 1998). The benefit of the meta-level method, especially for the content/collaborative hybrid, is that the learned model is a compressed representation of a user’s interest, and a collaborative mechanism that follows can operate on this information-dense representation more easily than on raw rating data.

In conclusion, hybridization can alleviate some of the problems associated with collaborative filtering and other recommendation techniques. Content/collaborative hybrids, regardless of type, will always demonstrate the ramp-up problem since both

techniques need a database of ratings. Still, such hybrids are popular, because in many situations such ratings already exist or can be inferred from data. Meta-techniques avoid the problem of sparsity by compressing ratings over many examples into a model, which can be more easily compared across users. Knowledge-based and utility-based techniques seem to be good candidates for hybridization since they are not subject to ramp-up problems. Table 2-10 summarizes some of the most prominent research in hybrid recommender systems.

All existing recommender systems employ one or more of a handful of basic techniques: content-based, collaborative, demographic, utility-based and knowledge-based. A survey of these techniques shows that they have complementary advantages and disadvantages. This fact has provided incentive for research in hybrid recommender systems that combine techniques for improved performance. Much recent research has been dedicated to the exploration of various hybrids, including the six hybridization techniques discussed in this paper: weighted, mixed, switching, feature combination, feature augmentation, and meta-level.

#### ***2.3.3.6 Evaluation Frameworks for Supporting Evaluators***

In the past 15 years, several evaluation frameworks for adaptive systems have been developed. For example, the Easy D hub<sup>12</sup> and European Quality Observatory (EQO) evaluation frameworks; EQO includes a recommendation service for the quality approaches, which is a repository of evaluation studies of adaptive systems. Weibelzahl (2003) developed an evaluation framework that both categorized existing studies and offered a systematic approach for evaluations. The researcher applied a layered evaluation approach in designing and evaluating the framework. The main objectives specified for this framework include what had to be evaluated to guarantee the success of adaptive systems, and a grid that facilitated the specification of criteria and methods that were useful for the evaluation. The second objective was aimed at encouraging further evaluations (Weibelzahl, 2003).

Gupta et al. (2004) proposed an evaluation framework for Adaptive Hypermedia Systems (AHSs). They pointed out that, although a number of frameworks were used in the evaluation of AHSs, the ones that applied the layered evaluation approach had proved

---

<sup>12</sup> <http://www.easy-hub.org>

useful in identifying the exact cause of adaptation failures or other errors in the system. (Tarpin-Bernard et al., 2009) developed an online framework that helped to characterize different types of adaptive features by helping the evaluator fill in a simple form. The information provided is then processed to obtain a quantitative evaluation of three parameters, called global, semi-global and local adaptation degrees. Based on the results of the analysed studies in this domain(Pawlowski, 2003), most of the existing evaluations of adaptive recommender systems have used the layered approach, which focuses on system perspective rather than the end-user, and very few recommend education evaluation data (i.e. European Quality Observatory framework) (Pawlowski, 2003).

Paramythis et al. 2010 propose a framework that can be used to guide the “layered” evaluation of adaptive systems, and a set of formative methods that have been tailored or specially developed for the evaluation of adaptivity. The proposed framework unifies previous approaches in the literature and has already been used, in various guises, in recent research work. The researchers further presented several methods are related to the layers in the proposed framework and the stages in the development lifecycle of interactive systems (Paramythis et al., 2010).

Having looked at AEL systems, the next sections discuss current evaluation approaches that have been used during evaluations of adaptive systems. This research focuses more specifically on the user-centred evaluation approaches.

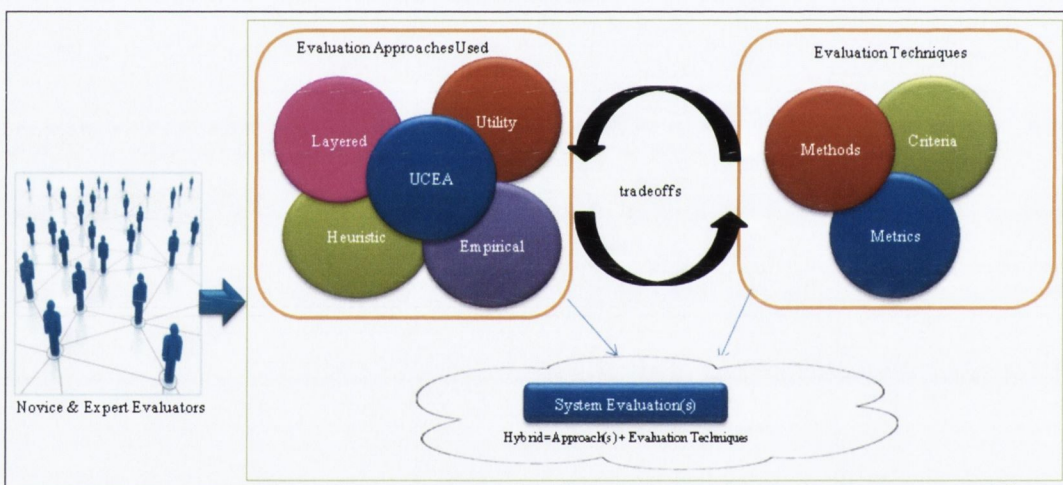
## **2.4 Comparison across Evaluation Approaches**

A comparison of evaluation approaches was conducted recently by evaluators of adaptive E-Learning systems (Mulwa C. et al., 2012). Several evaluation approaches were used in evaluating AEL systems (Vavoula and Sharples, 2009). In this section the candidate presents an overview of evaluation approaches used and discusses the tradeoffs between them. The aim is to address the issue that it is difficult to see what your evaluation approach should be in order to meet your evaluation objective, and what should be your range of evaluation techniques.

In order to produce effective results, evaluation should occur throughout the entire design cycle and provide feedback for design modification (Gena and Weibelzahl, 2007). 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 that have the purpose of evaluating the quality of learning resources. Figure 2-8 demonstrates the cross-over between evaluation approaches and evaluation techniques (methods, metrics and criteria).

Vourikari et al. (2008) conducted a review of evaluation approaches for learning resources. The researchers performed a tentative classification of these approaches and discovered a plethora of evaluation approaches for digital learning resources (Vuorikari 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 own quality requirement. Furthermore, Manouselis and Costopoulou (2006) acknowledge the diverse evaluation approaches for learning resources (such as models, methods, criteria and instruments) that are applied to ensure the quality of the learning resources (Manouselis and Costopoulou, 2006).



**Figure 2-8: Cross-over of evaluation approaches and techniques**

### 2.4.1 User-Centred Evaluation Approach

The *user-centred approach* is central to the creation of usable adaptive information systems (AISs) and adaptive educational E-Learning systems, services and institutions. A user-centred evaluation approach is prominent in many interface studies and has been proved effective during evaluations (Mulwa C. et al., 2011, Van Velsen et al., 2008, Mulwa C. et al., 2010). The benefits of the user-centred approach are savings in time and

cost, ensuring the completeness of system functionality, minimizing the required repair efforts, and improving user satisfaction (Nielsen, 1993). A discussion of this approach and its benefits was presented in Section 2.2.2.

## **2.4.2 Layered Evaluation Approach**

The *layered evaluation approach* (Karagiannidis C. and Sampson D., 2000, Brusilovsky P. et al., 2001) separates the 'interaction assessment' and the 'adaptation decision'. Both layers should be evaluated separately in order to effectively interpret the evaluation results. Evaluating an AEL system on a layer by layer basis has been recommended as a more comprehensive approach (Brusilovsky P. et al., 2001, Weibelzahl and Weber, 2002, Paramythis A. et al., 2010).

In contrast to other approaches that focus on the overall user's performance and satisfaction (Chin, 2001), layered evaluation in particular assesses the success of adaptation by decomposing it into different layers and evaluating each layer individually. 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, generalization of evaluation results, and re-use of successful practices.

## **2.4.2 Empirical Evaluation Approach**

The *empirical approach* helps 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. Empirical evaluations, also known as controlled experiments, refer to the appraisal of a theory by observation in experiments. 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).

## **2.4.3 Utility-Based Evaluation Approach**

Current evaluation practices attempt to evaluate adaptation as a whole, with user satisfaction or performance as the overall metric for success, based on identified measurable criteria. In the *utility-based 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. For example, if one compares an adaptive system with its non-adaptive counterpart, the value of adaptation is the difference in utility between the two systems. As described above, the main advantage of layered evaluation methods is that they break the utility function into several distinct functions. For example, suppose there is a utility  $X_1$  that maps the interaction assessment and the resulting user model to a real number that represents its correctness; suppose there is also a utility function  $U_2$  that maps a system, given some user model, to a real number that represents user satisfaction or performance. In this case the whole utility function can be expressed as  $X = X_1 X_2$ . It is clear that the latter utility function better indicates the usability of an adaptive hypermedia system. Utility-based evaluation of adaptive systems (Herder E., 2003, Mulwa. et al., 2011) offers a perspective on how to reintegrate the different layers.

#### **2.4.4 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. The heuristic evaluation technique is the most widely used inspection method. Heuristic evaluation uses a small set of evaluators who judge a user interface for compliance with usability design principles (Scholtz, 2004). A heuristic evaluation approach that 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 (Fu et al., 2002). This approach can help evaluators by improving the detection and diagnosis of potential usability problems.

Scholtz (2004), states that heuristic reviews are less expensive and less time-consuming to conduct than user-centred evaluations (Scholtz, 2004). The cognitive walkthrough can be accomplished using only a text description of the user interface and therefore can be used very early in the software development process. Inspection techniques do not provide possible solutions to the usability problem. Moreover, it is difficult to summarize the findings from multiple evaluators as they report problems differently and at different levels (Scholtz, 2004). There is also the issue of severity. Not all usability problems are equal. Development teams need to be able to prioritize which problems get fixed according to the seriousness of the problem. There is currently no agreement on how to judge the severity of usability problems.

### **2.4.5 Quality, Lifecycle & Combined Four-level and Six-level Approach**

The quality approach is used to investigate the current state of E-Learning quality in Europe. It is based on a survey by the European Quality Observatory (EQO), the European platform for quality in E-Learning involving 1,700 participants from all European countries (Ehlers et al., 2005). The combined *four-level approach* (Kirkpatrick Schenkel) and *six-level approach* (Breitner and Hoppe 2005) focuses mainly on pedagogical objectives. The *lifecycle approach* to educational technology evaluation 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 methods and ideas from software engineering educational evaluation and models for evaluating learning.

Having looked at different evaluation approaches and identified different evaluation techniques (methods, metrics and criteria) and mapped those techniques to approaches. After extensive review of literature, the most common used approach was the layered approach, followed by user-centred and empirical approaches respectively.

### **2.4.6 Design Decisions: Critique of Evaluation Approaches**

During our analysis of approaches used by both novice and expert evaluators, it was difficult to identify the user's evaluation objective, range of evaluation approaches and range of evaluation choices. In an evidence-based real-life study, The candidate looked through what evaluators from three communities (adaptive hypermedia, user modelling, adaptation and personalization and recommender systems) had evaluated for their adaptive systems developed from 2000 to 2012, and what techniques they had used, and mapped those techniques to different methods, criteria and metrics. It is clear that these researchers need more advice around their evaluation options in order to reach their goal. The candidate wanted to show that there are many design choices and decisions made between having an adaptive system, defining your objective as to what you want to evaluate, and the actual evaluation technique (method, criteria and metric). Table 2-11 presents an example of design decisions made concerning which evaluation approaches to use and other approaches they could have used but opted not to.

**Table 2- 11: Design decisions made on which evaluation approaches to use and other approaches that could have been used**

<b>System Name</b>	<b>Evaluation Objective</b>	<b>Approaches Possible</b>	<b>Approach Used</b>	<b>Candidate Approaches</b>
APeLS-Activity Based Personalized E-Learning	To determine the usability of the personalized SQL course, in particular focusing on learner satisfaction and the effectiveness of the service.	Any of the above depending on evaluation objective.	User-centred evaluation approach.	Layered & utility evaluation approach.
PEACH	To assess the attitudes towards the four adaptivity dimensions, through two simulated video museum guides, an adaptive one (AD) and a non-adaptive (NAD).	Any of the above depending on evaluation objective.	Empirical	User-centred evaluation approach & heuristic.
ERM-Tutor (2005)	To investigate the usage of free-form questions.	Any of the above depending on evaluation objective.	*Not reported	User-centred evaluation approach.

**Table 3 Continued**

MastroCARONTE	Tested whether the system can provide immediately the pieces of information that are most suitable for a user, presenting them in a way which is compatible with the user capabilities and the risk of the contextual situation.	Any of the above depending on evaluation objective.	*Not reported	A hybrid of User-centered and layered evaluation approach.
ARCHING-Adaptive Information Retrieval and Composition System	Testing and evaluating the adaptive composition architecture with real users to find out educational benefits and user satisfaction of the service.	Any of the above depending on evaluation objective	*Not reported	A hybrid of user-centred and empirical evaluation approach.

Having looked at this data presented in sections 2.4.1 to 2.4.7, the question is asked ‘what did people choose to do versus what they could have done?’ This shows that, although people did one or more things, there were lots of other options they could have chosen during evaluation but did not. There are many design decisions to be taken ‘along the way’; for expert evaluators this is tricky enough but for a novice evaluator it’s more difficult. In the published papers, people do not give the reasons for many of their evaluations (i.e. this is appropriate for this). Therefore, there is a need to provide greater support in decision-making, to build up a knowledge base to aid evaluation choices and techniques to use.

## 2.5 Tradeoffs between Techniques to Support User-Centred Evaluations

### 2.5.1 Summary and Critique

The methodologies for evaluating adaptive E-Learning systems are generally borrowed from the methodologies used in human computer interaction (HCI) and those used for the evaluation of the information selection process (Gena, 2005). In this work, evaluation techniques are the methods, criteria and metrics. The HCI methodologies can be used in the evaluation of AEL systems mostly to evaluate the interface adaptations, the usability of adaptive systems, to collect users' and experts' opinions, etc. Such methodologies can also be used for the evaluation of the information selection process in order to collect user data important to the analysis of the process. Evaluation of these systems is a significant but very complex area of research in itself, depending on the aspect of adaptivity that needs to be evaluated. Several evaluation techniques need to be combined and executed differently in order to produce good results. Figure 2-9 depicts tradeoffs between different evaluation techniques for AEL systems. Tables 2-12 and 2-13 present examples of these techniques used from 2000 to 2012.

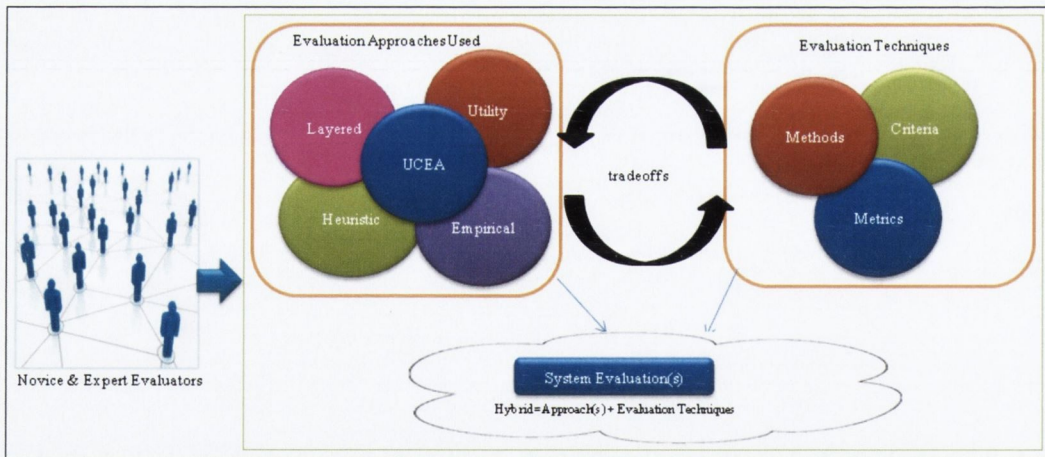


Figure 2-9: Tradeoffs between different evaluation techniques for AEL systems

**Table 2- 12: Tradeoffs between evaluation techniques (methods, criteria and metrics)**

<b>Evaluation Method/ Instrument</b>	<b>Criteria (Variables)</b>	<b>Metrics</b>	<b>References</b>
Interviews, questionnaires (online, post-test, pre-post-test, verbal), focus group, discussion groups.	Usability, perceived usefulness, intention to use, user goals, knowledge of the domain, background and hyperspace experience, preferences trust and privacy issues, appropriateness of adaptation.	Accuracy of recommendations, accuracy of retrieval, AiAI: administrator interaction adaptivity index.	(Gena, 2005c), (Van Velsen et al., 2008), (Masthoff, 2006, Raibulet and Masciadri, 2009) .
User observation, systematic observation, verbal protocol, data mining, play with layer, simulated users, cross-validation, heuristic evaluation, play with layer, simulated users, cross validation	Usability, User behaviour, user goal, knowledge of domain, background and hyperspace experience, user interests individual traits (e.g. cognitive or learning style), environment (e.g. location, locale, software, hardware), user situation awareness.	Behavioural complexity, reliability metrics, precision, software size and length metrics, UiAI: User interaction Adaptivity index.	(Gupta and Grover, 2004), Rothock et al. 2002, (Magoulas and Dimakopoulos, 2005), Steehouder M. 2008, (Brusilovsky, 2001 )
Heuristic evaluation, expert review, parallel design, cognitive walkthroughs, social-technical models	Usability of interface adaptation & user, domain and interface knowledge, user performance.	pQoR: performance Quality of Response.	Rothock et al. 2002
Wizard of Oz simulation, scenario-based design, prototypes	Early prototype evaluations, evaluation before implementation.	pIA: performance Influence on Adaptivity.	Judith Masthoff 2006

Table 2- 12 Continued

Evaluation Method/ Instrument	Criteria (Variables)	Metrics	References
Usability testing, experimental evaluation	Interface (and content) adaptation, usage data (user history), user cognitive workload, groups of users.	MpAC: Minimum personalisation adaptive cost.	Magoulas & Demakopoulos, Rothock et al. 2002.
Cultural probes, focus group, user-as-wizard, heuristic evaluation, cognitive walk through, simulated users, play with layer, user test.	Preferences, user interests, user skills and capabilities, user performance.	AvgpACF: Average personalisation adaptive cost per functionality.	(Santos, 2008), (Masthoff, 2006), (Paramythis et al., 2010)
Creative brainstorming sessions, focus group, user-as-wizard.	Privacy, transparency, appropriateness, appreciation, trust and privacy issues, user experience, user satisfaction, usability, user behaviour, intention to use, perceived usefulness.	MpOCF: Minimum personalisation, Overall Cost.	(Van Velsen et al., 2008)
Quantitative, grounded theory, cognitive walkthrough, heuristic evaluation, user test	To combine qualitative evaluation, to discover new theories.	ApOC: Adaptive personalisation Overall Cost.	Diaz et al. 2008, Gena 2005.
Prototyping, heuristic evaluation, cognitive walkthrough, user test, play with layer, cooperative evaluation, verbal protocols, and focus group.	Evaluation of vertical or horizontal prototype, Collaboration with real users during the final evaluation step.	DSAI: Domain specific Adaptivity index.	Gena 2005

### **2.5.1.1 Evaluation Methods**

The methodologies for evaluating adaptive recommender systems are generally borrowed from the methodologies used in human computer interaction (HCI) and those used for the evaluation of the information selection process (Gena, 2005). In the analysed studies, questionnaires, experimental evaluation, interviews, user observations, usability testing and user texts were the most commonly used evaluation methods. *Questionnaires* are used to collect data from respondents by allowing them to answer a set of questions either on paper or online. Participants can choose one or multiple choices or can freely answer in writing.

It is essential to not only evaluate but also to ensure that the evaluation uses the correct methods since an incorrect method can lead to wrong conclusions (Gena and Weibelzahl, 2007, De Jong and Schellens, 1997)

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, as well as interviews, user observations and usability testing. In interviews (i.e. structured, fixed questions, or semi-structured), participants normally are questioned in person by an interviewer. The manner in which interview results were reported indicates that evaluators considered 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. This focuses on measuring the system's capacity to meet its intended purpose. In total, 40 evaluation methods were mentioned in the studies.

### **2.5.1.2 Measurement Criteria**

The evaluation of AEL 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). 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. *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 categories of variables concerning attitude and experience, actual use, system



adoption and system output. *Usability* was most frequently measured, followed by *User Satisfaction*, a subjective variable which can be influenced by factors such as 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: (i) the product's design process, (ii) the product itself, (iii) use of product and (iv) user experience of the product and user expectation (Nokelainen, 2006).

The candidate has identified adaptive variables (also known as concepts) that can prompt adaptivity, in the literature from 1996 to 2008. These variables make AEHS a variable tool for learners in TELE. A total of 21 adaptive variables that can prompt adaptivity were identified (Mulwa C. et al., 2010).

Brusilovsky (1996) identified the following features as currently used by adaptive hypermedia systems: i.e., user's goals, knowledge, background and hyperspace, experience, and preferences. The researcher in 2001 added two more variables: user's interest and individual traits. 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 personalized information access. Velsen et al (2008) identified 13 variables concerning UCE and grouped them in the following categories: (i) variables concerning attitude and experience (i.e., appreciation, trust and privacy issues, user experience and user satisfaction, (ii) variables concerning actual use (i.e., usability, user behaviour and user performance, (iii) variables concerning system adoption (intention to use, perceived usefulness) and (iv) variables concerning system output (appropriateness of adaptation, comprehensibility and unobtrusiveness) (Van Velsen et al., 2008). The researchers provide a list of how often each variable was addressed in the 63 studies they reviewed and accept that the wording of most variables spoke for itself. These variables are very significant since the scope of the user model of the architectural model for the evaluation module is based on them.

### **2.5.1.3 Evaluation Metrics**

Many metrics can be used to measure performance; for example, knowledge gain, amount of requested materials, duration of interaction, and number of navigation steps, task success, usability (e.g. effectiveness, efficiency and user satisfaction). In the analysed studies, accuracy of recommendations metric was the most frequently used, followed by accuracy of retrieval.

As noted above, many design decisions are made by evaluators when it comes to which technique is most appropriate for their evaluation objectives. Section 2.5.2 presents a sample of these design decisions and other options they could have used but did not.

## 2.5.2 Design Decisions: Critique of Evaluation Techniques Used

Tables 2-13 and present a summary of decisions made by evaluators on which evaluation techniques (i.e. methods and criteria) to use and other techniques they could have used but opted not to.

**Table 2- 13: Examples of design decisions on evaluation methods**

<b>System Name</b>	<b>Evaluation Objective</b>	<b>Method Possible</b>	<b>Method Used</b>	<b>Candidate Methods</b>
APeLS-Activity Based Personalized E-Learning.	To determine the usability of the personalized SQL course, in particular focusing on learner satisfaction and the effectiveness of the service.	Any user-centred evaluation methodologies.	Paper-based questionnaire.	Task-based, usability testing, Expert Review.
PEACH	To assess the attitudes towards the four adaptivity dimensions, through two simulated video museum guides, an adaptive one (AD) and a non-adaptive (NAD).	Any user-centred evaluation & heuristic methodology.	Interviews, Questionnaires, Simulated Users, Experimental Evaluation, Empirical Observations.	Task-based, Quantative methods.

Table 2- 13 Continued

System Name	Evaluation Objective	Method Possible	Method Used	Candidate Methods
ERM-Tutor (2005)	To investigate the usage of free-form questions.	Task-based method, structured interviews.	Questionnaires, Focus Group, User Observation, Data Mining, Cross-Validation, Prototyping, Expert Review, Wizard of Oz Simulation, Usability Testing, Experimental Evaluation, User Test, Creative Brainstorming Sessions, Empirical Observations.	Any of user-centred evaluation methods.
MastroCA RONTE	Tested whether the system can provide immediately the pieces of information that are most suitable for a user, presenting them in a way which is compatible with the user capabilities and the risk of the contextual situation.	usability testing methods	Pre & post questionnaire, logs recording.	

**Table 2- 14: Examples of design decisions on evaluation criteria**

<b>System Name</b>	<b>Evaluation Objective</b>	<b>Criteria Possible</b>	<b>Criteria Used</b>	<b>Candidate Criteria</b>
APeLS-Activity Based Personalized E-Learning	To determine the usability of the personalized SQL course, in particular focusing on learner satisfaction and the effectiveness of the service	Learnability, effectiveness	Usability, Perceived Usefulness, Appropriateness of Adaptation, User Behaviour, User Satisfaction,	Content Adaptation, Preferences, User Skills and Capabilities, User Performance
PEACH	To assess the attitudes towards the four adaptivity dimensions, through two simulated video museum guides, an adaptive one (AD) and a non-adaptive (NAD).	Early prototype Evaluations	Usability, Perceived Usefulness, Intention to Use, User Behaviour.	User Satisfaction.
ERM-Tutor (2005)	To investigate the usage of free-form questions.	Real User Actions.	Usability, Perceived Usefulness, User Behaviour, Content Adaptation, Preferences.	User Performance, User Cognitive Workload.
MastroCARONTE	Tested whether the system can provide immediately the pieces of information that are most suitable for a user, presenting them in a way which is compatible with the user capabilities and the risk of the contextual situation.	Knowledge, User skills & Capabilities, Usability (Satisfaction), Interface.	Usability, Perceived Usefulness, Appropriateness of Adaptation, User Performance, Real User Actions.	User Behaviour, User Goal, Knowledge of Domain, Background and Hyperspace Experience.

Having looked at all this data (of evaluation techniques), the question is asked: *“What did people choose to do verse what they should have done?”* Although evaluators did one or two things; there were lots of other options they could have chosen and did not, some of which might have been more difficult than others.

It is clear that there are a lot of design decisions along the way. People will make their design decisions, the difficulty is for an expert evaluator it is tricky enough but for a novice evaluator it is even more difficult. Therefore there is a need to provide support in the decision making. Furthermore there are a lot of choices being made, but in the papers people do not give reasons. It would be good to build a knowledge base of evidence based approach to deciding evaluation choices and techniques.

## **2.6 Challenges, Problems, Issues and Limitations during Evaluations**

Adaptive technology E-Learning has attracted much interest with its 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 AEL system reacts differently for each individual user and each context of use. In the evaluation of adaptive systems, difficulties can be caused by the need to distinguish different adaptation aspects and therefore evaluate them separately.

### **2.6.1 Critique of Challenges**

A key challenge faced by evaluators of AEL systems is the difficulty in choosing the right evaluation approach and technique to use. Furthermore, the evaluation of the overall performance of such systems is a major challenge, as the AEL system reacts differently for each individual user and context of use. The evaluation of such systems (presented in Section 2.3.3) is a difficult task (Lawless et al., 2010, Tintarev and Masthoff, 2009). Several researchers have emphasized the difficulties caused by the complexity of such systems and the usability issues raised by end users (Missier Del and Ricci, 2003, Lavie et al., 2005, Weibelzahl and Weber, 2002, Markham et al., 2003). Other challenges include identifying the appropriate evaluation educational datasets (discussed in Section 7.3).

A key challenge faced by evaluators of AEL systems is the difficulty in choosing the right evaluation approach and technique to use. The evaluation of adaptive TEL systems such as adaptive recommender systems is a difficult task (Lawless et al., 2010, Tintarev and Masthoff, 2009).

## **2.6.2 Critique of Problems and Issues**

Overall, the candidate believes that 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.

One major problem lies in understanding the adaptation mechanism of the system; what is improved by the adaptation and what might have been the situation if a different kind of adaptation had applied. It is difficult to define the effectiveness of adaptation. When users work with an adaptive system, it is very difficult in principle to demonstrate what 'might have been' or what impact the system's adaptive processes actually had on the end-user.

- The preventive measures' aim is to ensure that a property is not present in such a manner that it would cause problems.
- The compensatory measures goal is to ensure that, in some other way, the goals and objectives are achieved despite the threats created by the properties challenges.

Other issues include pitfalls encountered by developers of these systems (Mulwa et al., 2010, Tintarev and Masthoff, 2009, Gena and Weibelzahl, 2007) such as:

- Difficulty in attributing cause: Is the adaptation causing the measured effect or another aspect of system functionality or design (e.g. system usability)?
- 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. 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.
- The selection of the metrics to be used in the evaluation of AHS is crucial. There are currently no agreed evaluation methodology standards, thus making AH evaluation a difficult, complex and time-consuming task.

## 2.7 Conclusions

To conclude, the candidate advocates the importance of evaluation of AEL systems. Significant evaluation results can lead to more appropriate and successful systems; their user's point of view can be a very inspiring source of information adaptation strategies. Since evaluation of adaptive systems, especially AEL systems, is still in the exploratory phase, new approaches are strongly called for. These can include combining different techniques, exploring new metrics to assess adaptivity, and adapting the evaluation technique to the adaptive system features.

The candidate believe this is the first evidence-based study of what people are doing in evaluations of adaptive systems and providing a real picture of who is doing what to evaluate what. Regarding this evidence-based approach, the question is now asked: *How can I capture it, manage it and allow it to be used easily?* Chapter 3 and 4 address this question.

## **Chapter 3: Overall Research Methodology & EFEx and OSSES System Architecture Design**

### **3.1 Introduction**

As outlined in chapter 2, the evaluation of adaptive systems is significance. Furthermore the evaluation of adaptive E-Learning systems helps improve the performance of such systems and increase better learning experience. This chapter briefly presents the influences from the state of the art. Section 3.2 presents a high level overview of the overall methodology used when conducting this research; Based on these influences and identified gaps; section 3.3 and 3.4 presents a high-level overview of the architectural design for the proposed evaluation framework for supporting novice and expert evaluators of adaptive systems (EFEx) and the proposed focused crawling system used to crawl general evaluation studies (OSSES). The proposed tools are significant since they can help support both novice and expert evaluators of such systems during the evaluation of design, implementation and evaluation phases of such systems.

### **3.2 Influences from State of the Art**

The aim of this section is to provide the reader with a summary of influences from the state of the art and how they affect the core components of EFEx evaluation framework and OSSES system. The proposed applications supports evaluator during design decision making between having an adaptive system and an objective of what they want to evaluate to the actual evaluation technique.

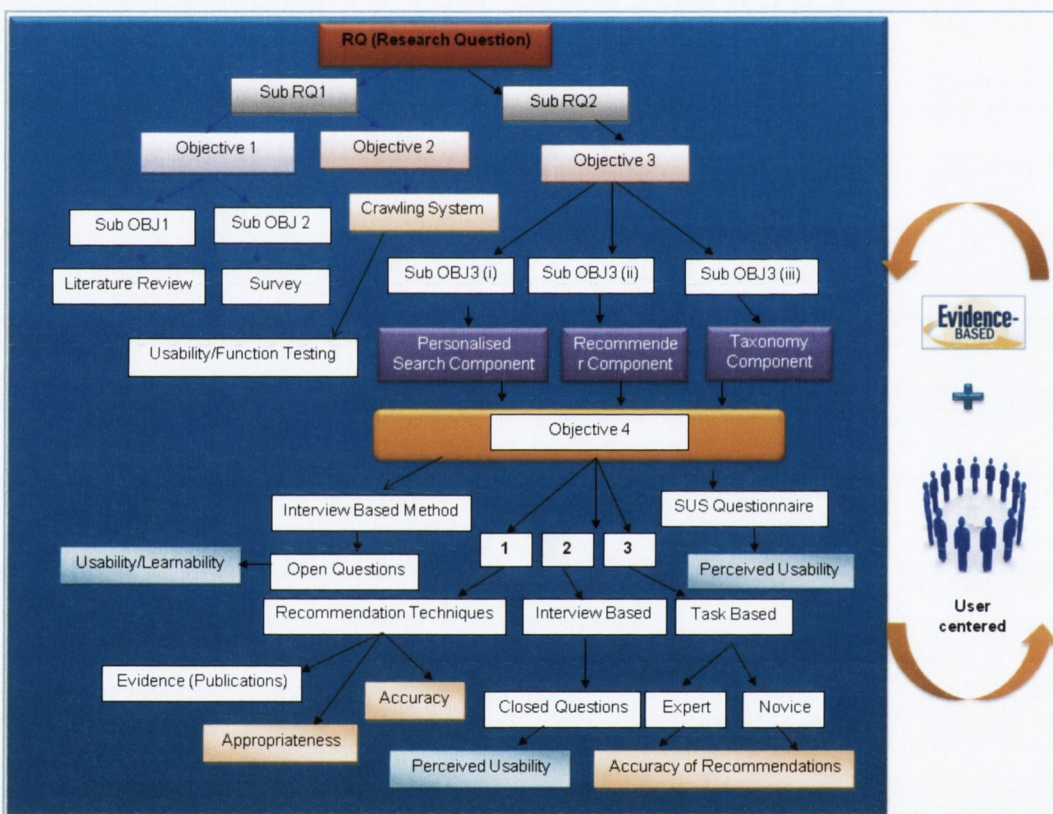
The analysis conducted in chapter 2 influenced various aspects of the outcome of this thesis. Furthermore the hybrid (evidence-based and user-centred) research approach used to conduct this research proved to be very effective. Based on the analysed results of chapter 2, which provided us with a clear picture example of; people using certain evaluation approaches and evaluation techniques (methods, metrics and criteria) to evaluate adaptive systems what the candidate is pointing out is that there are lots of design



choices and therefore there is need to provide better information about “**What is theoretically appropriate and what people are doing**”. It is now asked:

- How do we identify theoretically what needs mapping?
- How do we identify what techniques people are using and the tradeoffs between those techniques to support user-centred evaluations of adaptive systems?

In order to answer these questions, there is need to design an evaluation framework for supporting novice and expert evaluators of adaptive systems and a focused crawling system. Section 3.3 and 3.4 presents detailed descriptions of requirement specification and design of these support tools which are major contributions of this research.



**Figure 3- 1: A Hybrid (evidence-based and user-Centred) research methodology**

### 3.3 Overall Methodology

This research applies a hybrid methodology consisting of evidence-based and user-centred (Van Velsen, et al., 2008). A methodology is defined as the systematic, theoretical analysis of the methods applied to the research field of adaptive systems. It is the theoretical analysis of the body of methods and principles associated with a branch of knowledge which typically, encompasses concepts such as paradigm, theoretic model, phases and quantitative or qualitative techniques.

The user-centred methodological approach (refer section 2.2.2) contribute to innovations in engineering design and have been shown to increase productivity, improve quality, reduce errors, improve acceptance of EEx evaluation framework and OSSES system and also reduce development costs. For this approach to be effective the candidate used different techniques (structured-interviews, quantitative close-ended questionnaires and task-based). Interviews helped us collect self-reported experience, opinion and behavioural motivations of both novice and expert evaluators of such systems. They were essential in finding out procedural knowledge as well as problems with the design of both applications. The online qualitative close-ended questionnaires

Furthermore the evidence-based approach enabled us to get a real picture of who is doing what to evaluate what. This evidence was collected from five communities (user modelling adaptation and personalisation (UMAP), recommender community, data Technology-enhanced (dataTEL), adaptive hypermedia and information retrieval). This approach proved to be very effective. During interaction with members of these communities, structured-interview based methods were used. Participants were asked questions such as:

- How they conducted adaptive systems belonging to different variation types (discussed in section 2.3.2)?
- Which properties they focused on and what was improved by adaptivity?
- What was the goal (purpose) of their evaluation?
- What aspects they had focused on?
- What questions they asked?
- Finally what kind of results they expected after evaluations?

## 3.4 Overall Architecture

This section introduces the overall architectural design of the major and minor contributions of this research; a high level overview of proposed evaluation framework for supporting novice and expert evaluators of adaptive systems (EFEx) and a focused crawling system for evaluation studies for evaluation studies in adaptive systems (OSSES).

### 3.4.1 EFEx - Evaluation Framework for Supporting Evaluators of Adaptive Systems

The architectural of EFEx 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. The architecture has a web-based interactive and collaborative interface consisting of the *presentation layer* which is the topmost level of the application which displays information related to services. The *business logic layer* which is pulled out from the presentation tier and, has its own layer, controls the frameworks' functionality by performing detailed processing. Finally the *data persistence layer* 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. The framework consists of three major components: a hybrid recommender system, a personalized search systems and taxonomy of technical terms. A high level architectural design of these components is depicted in Figure 3-2.

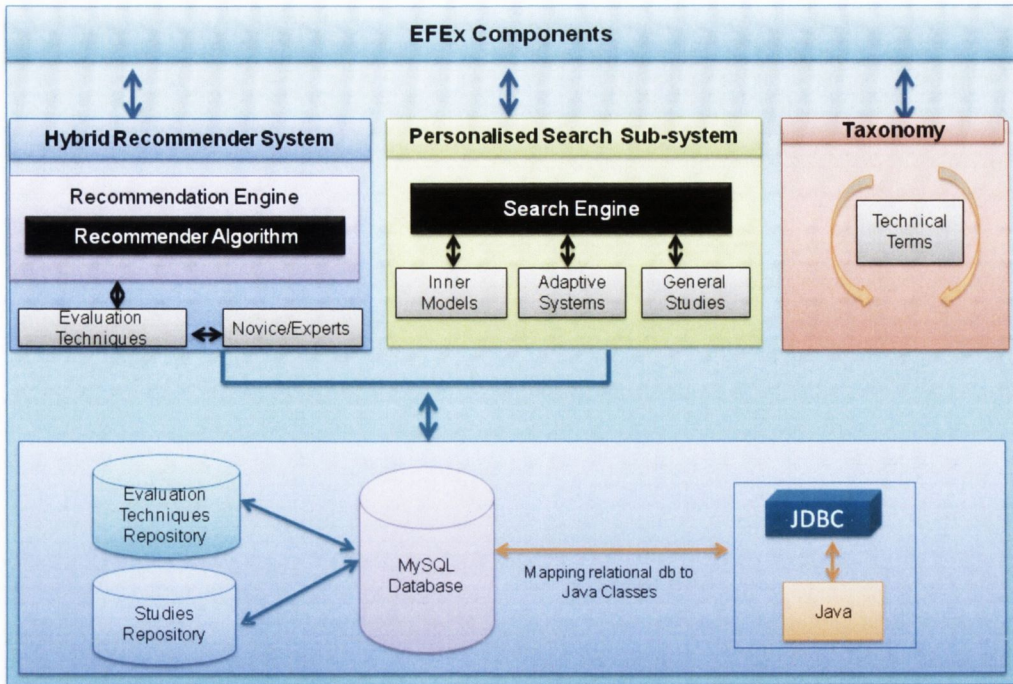


Figure 3- 2: High Level Overview of EEx Framework Architectural Design

Following is a brief description of each component:

1. A novel *recommender system* for evaluating adaptive E-Learning systems built upon an evaluation educational dataset using a hybrid (case-based and knowledge-based) evaluation methodology to identify appropriate evaluation techniques. This methodology overcomes the limitations of case-based and knowledge-based methods (discussed in section 2.3.3.5). Recommendation technology is used in order to enhance the appropriateness of suggestions of evaluation techniques for adaptive E-Learning systems. In particular the multi attribute relationships which need to be traversed by humans to work out what are the most appropriate evaluation techniques are not easily navigated using typical database techniques. The database is populated using an educational evaluation dataset that consists of a characterised, structured and interlinked list of (9 evaluation approaches, 84 methods, 85 metrics, 74 criteria (also known as adaptive factors), 15 metadata internal models of adaptive systems, 106 adaptive systems (developed from 2000 to 2013) and 16 variations types of these systems.
2. A *Personalized Search System* whose database is populated using 250 evaluation studies (2000 to 2013) of adaptive systems. The users are provided with an automated web-based personalized search interface, which is divided into three user interfaces which allow users find: evaluation studies of the internal models of

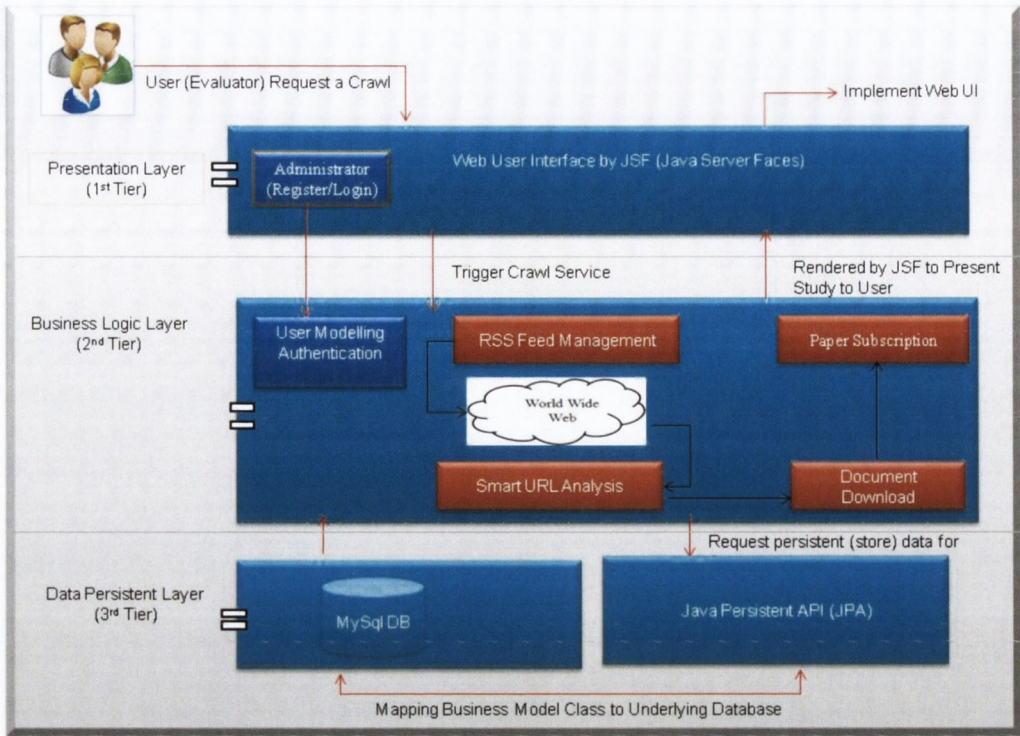
such systems; evaluation studies of adaptive systems developed from 2000 to 2013 and finally general evaluation studies of such system.

3. A *Taxonomy of Technical Terms* of evaluations of adaptive systems which assists novice evaluators in understanding different aspects of such systems.

These components are discussed in chapter 4. Finally the chapter presents the results and findings of the design and prototype testing respectively.

### **3.4.2 OSSES - Focused Online Crawling Systems for Evaluation Studies**

In recent years, due to exponential increase in the number of internet users, finding the appropriate information in the World Wide Web (www) is difficult and a challenging task. Web crawlers represent a significant component in Web search engines. The main educational benefit of the proposed crawling system is to provide a reference tool that has an interactive database to encourage evaluations of systems that fulfil certain methodological requirements. The synopsis of studies collected can be used as a basis of a searchable online database that provides an overview of the state-of-the-art to the scientific community and encourages other scientists to evaluate their own system. The system will support researchers to identify pitfalls in the planning process as well as in the analysis of collected data and also identify omissions in the state-of-the-art in future.



**Figure 3- 3: High-level Architectural Design of OSSES Crawling Systems**

The collaborative nature of the system enables sharing information among research students providing them with a larger view of the state-of-the-art. The system is designed using a three tier architecture design which is composed of rich site summary (RSS) Feed Management, Paper Subscription, Smart URL Analysis and Document Downloading. The RSS Feed Management allows a user to manage a set of Web feed formats that will publish most recent papers. As soon as a paper is published via RSS Feed, the paper subscription module automatically creates metadata. Upon receiving the document link, the Document Downloading module copies the document to a local repository (Mulwa C. et al., 2010). Figure 3-3 depicts the overall architecture of the crawling system.

### 3.5 Conclusion

The aim of this chapter was to present a high level overview of the methodology used when conducting this research and a high-level overview of the architectural design EFEx evaluation framework and OSSES system components. Chapters 4 and 5 discuss the architectural, technical design and implementation of both EFEx framework components and OSSES crawling system. The chapters also present the results of design and prototype testing.

# **Chapter 4: Evaluation Framework for Supporting Evaluators of Adaptive Systems**

## **4.1 Introduction**

As outlined in Chapters 1 and 2, adaptive E-Learning (AEL) systems typically are capable of providing appropriate information to the right learner at the right time, while keeping track of usage and also accommodating content dynamically to each learner and for the best learning results. However, evaluators of such systems are faced with challenges due to usability issues and the complexity of such systems. This chapter describes the design and implementation of the proposed evaluation framework (see Chapter 3) for supporting novice and expert evaluators of adaptive systems (EFEx).<sup>13</sup> The key objectives and scope of the proposed evaluation framework are also discussed.

The chapter is structured as follows. Section 4.3 describes the main components of EFEx architecture and technical design, which comprises three major components: a hybrid (case-based and knowledge-based) recommendation system, a personalized search system that allows evaluators to find evaluation studies of adaptive systems, and a taxonomy of technical terms for supporting the evaluation of adaptive systems. The section also presents design testing results. Section 4.4 discusses the implementation of the three components introduced in Section 4.3 and presents the results of testing the developed prototype.

## **4.2 Objectives and Scope of EFEx Evaluation Framework**

In developing an evaluation framework for supporting novice and expert evaluators of adaptive systems (EFEx), the candidate pursued four objectives:

- To specify what has to be evaluated to guarantee the success of adaptive systems.
- To establish an automated service for recommending appropriate evaluation techniques (methods, metrics and criteria) and evaluation approaches to be used.
- To provide a personalized search knowledge base that stores evaluation studies of adaptive systems.

---

<sup>13</sup> A framework refers to several instances (systems and taxonomy) running under the same platform.

- To provide a taxonomy of technical terms for adaptive systems

Regarding the first goal, currently there are multiple design decisions (see Section 2.5.3) being made about which evaluation techniques and approaches to use. For novice evaluators, this is particularly difficult. A framework like the one proposed here can help to systemize current approaches and also support evaluators of such systems by recommending which evaluation techniques to use, and providing explanations as to why that technique was recommended.

The second goal is important in that research has shown that evaluators of adaptive systems encounter many challenges and problems (discussed in Section 2.6). Provision of a recommendation service that is built on an evaluation educational dataset, using a hybrid (case-based and knowledge-based) evaluation method to identify appropriate evaluation techniques, will facilitate the design decision-making process both before and during evaluations of such systems.

Achieving the third goal is important to encourage further evaluations through providing a centralized knowledge base that supports and helps users find evaluation studies of internal models and adaptive systems. The final goal is important especially in relation to users who are new to evaluation and do not understand some of the technical terms used.

### **4.3 Architectural Design and Technological Design**

This section discusses the main components of EFEx architectural and technical designs. Design is defined as: the systematic, intelligent process in which the candidate use to generate, evaluate and specify concepts for the EFEx framework and the processes whose form and function achieve the adaptive E-Learning evaluator's objectives or users' needs while satisfying a specified set of constraints. This definition is partly based on the design definition of (Dym et al., 2005).



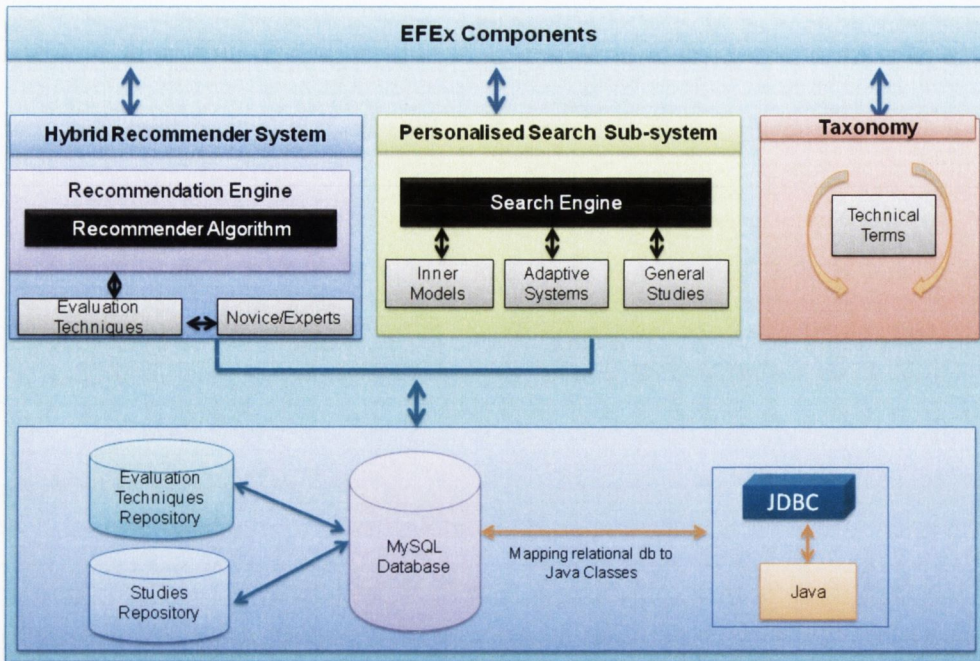


Figure 4- 1: High Level Overview of EEx Framework Architectural Design

### 4.3.1 Architecture Components and Capabilities

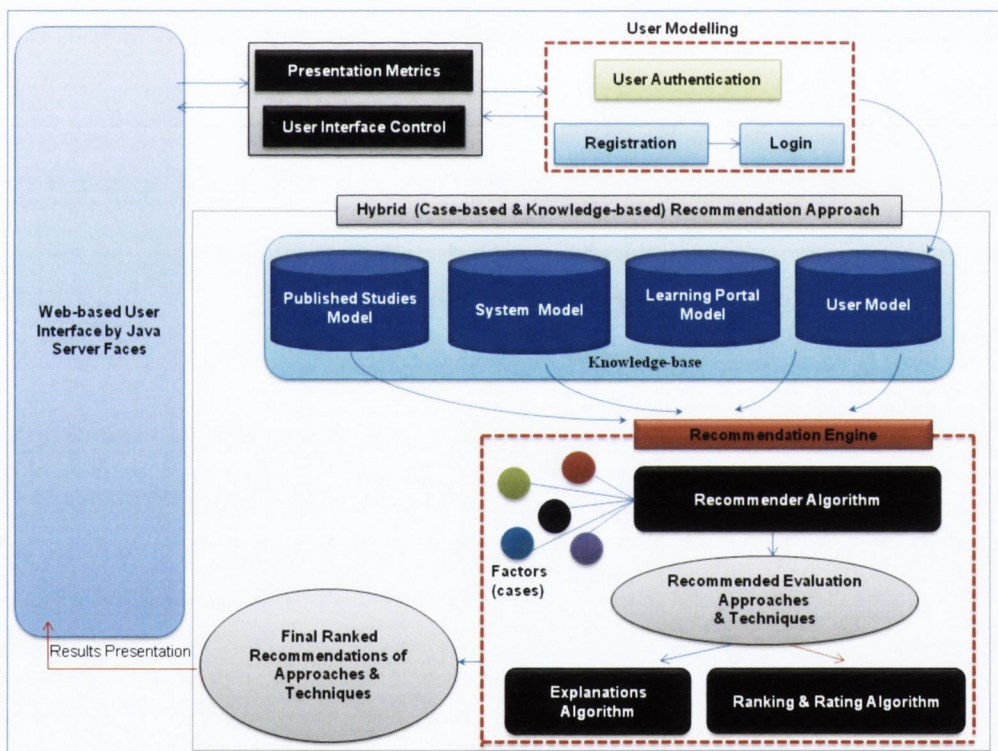
The recommendation system consists of four components and is designed as a web-based three-tier architecture (depicted in Figure 4-2): the *presentation layer*, which displays recommended evaluation approaches, evaluation techniques and evaluation bundles (method, criteria and metric) to the end user; the *business logic layer*, which is pulled out from the presentation tier, and controls the recommendation component functionality by performing detailed processing, and the *data persistence layer*, which keeps data neutral and independent from application servers or business logic.

#### 4.3.1.1 An Automated Novel Hybrid Automated Recommendation System

Recommendation techniques have a number of possible classifications (Burke, 2002). Such systems have: (i) background data – the information that the system has before the recommendation process begins; (ii) input data – the information that the user must communicate to the system in order to generate a recommendation, and (iii) an algorithm that combines the background and input data to arrive at suggestions.

This section discusses the architecture of a hybrid (case-based and knowledge-based) recommender system, one of the components of the EFEx evaluation framework. A review of current recommendation techniques (Mulwa. C., et al., 2012) proved that hybrid recommender systems combine two or more recommendation techniques in order to achieve better performance, with fewer of the drawbacks of any individual one.

The candidate proposes a hybrid (case-based and knowledge-based) recommendation service for supporting novice and expert evaluators of adaptive systems, built on an evaluation educational dataset using a hybrid (case-based and knowledge-based) evaluation method to identify appropriate evaluation techniques. This evaluation method overcomes the limitations of case-based and knowledge-based methods. The hybrid recommendation service is built on an educational evaluation dataset (discussed in Chapter 5).



**Figure 4-2: Overview of the various components of the hybrid recommendation service**

The case-based reasoning (CBR) technique was chosen because it is a problem-solver that uses the recall of examples as the fundamental problem-solving process. A case-based recommendation service is one that treats the objects to be recommended as cases, and uses CBR techniques to locate them. It contains a number of different “knowledge

containers” (Burke, 2002): the *case base*; the *vocabulary* in which cases are described; the similarity *measure* used to compare cases, and, if necessary, the knowledge needed to *transform* recalled solutions. In building a case-based system, the developer can choose where in the system different types of knowledge can reside. A low-level vocabulary for cases may push more complexity and hence more knowledge into the similarity measure. On the other hand, the knowledge-based recommendation technique attempts to suggest objects based on inferences about a user’s needs and preferences. This approach is distinguished by the fact that it involves functional knowledge: knowledge about how a particular item meets a particular user need. It is thus possible to reason about the relationship between a need and a possible recommendation.

The capability of four components of the recommender system is briefly discussed below:

### ***User Modelling***

User modelling enables us to customize and adapt the system to the novice and expert evaluator’s specific needs. A user model was used that stores details (name, password, email and organization). It allows a more up-to-date representation of evaluators who are logged into the system. Changes in their interests, their learning progress on what type of support they need, and interactions with the system are noticed, and influence the user model.

The model capabilities enable the creation of authentication components, which include registration before evaluators are allowed to start interacting with the system. Presentation metrics and user interface control mechanisms provide capabilities such as personalization to the user’s specific needs. The process of authentication involves confirming the identity of an evaluator in order to assist them when they encounter any usability issues. This is significant because some evaluators might have developed a new adaptive system; if they need recommendations on which evaluation approach to use or which evaluation technique to apply, they need to submit data (system name, characteristics and variation type) into the recommender system. Thus authentication (registration and login) enables the system administrator to keep track of who is submitting what data; it is hoped that this will help in maintaining data integrity and also in case a novice evaluator needs extra advice. An advantage of the dynamic user model is that it can be updated and can take into account the current needs and goals of the evaluator.

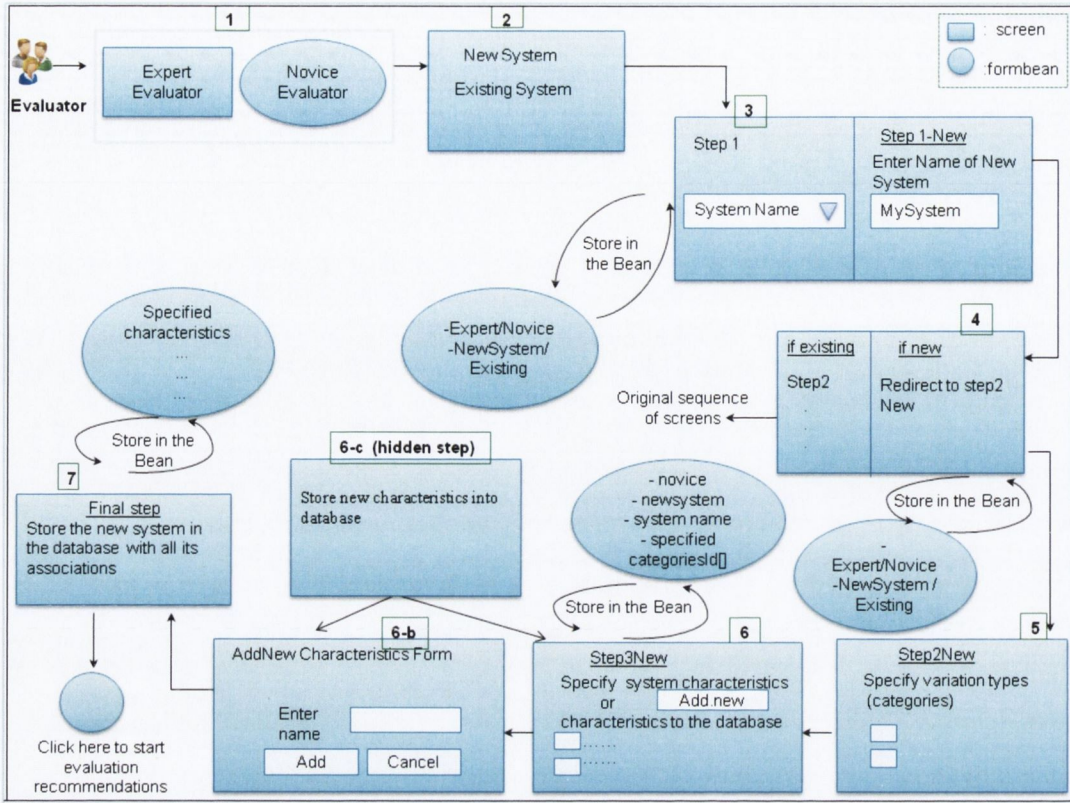
## ***Recommendation Engine***

The recommendation engine domain is large and can be very complex. There are some important decisions to make when researchers decide to start personalizing their system and want to use the recommendation engine. Recommendation engines are arguably one of the trendiest uses of data science. They can solve the problem of connecting researchers' existing users with the right items in their massive inventory of content. They can apply a variety of patterns and analyze user habits to offer recommendations to evaluators, and be helpful in presenting offerings that a user might not otherwise know about. The recommendation engine that is used has capabilities such as driving users to explore offerings from our system. The main purpose of a recommendation engine is to make inferences on existing data to show relationships between objects (users and items).

In this research the recommendation engine consists of two hybrid (case-based and knowledge-based) in-built algorithms: the recommender algorithm for evaluation (techniques and approaches), and the explanation algorithm, used to apply our educational data and also to analyze available information and provide personalized and real-time recommendations. The design of the recommendation engine depends on the domain (e.g. adaptive E-Learning) and the particular characteristics of the data.

## ***Recommender Algorithm***

The recommender algorithm has the capability to process the available information and provide real-time and personalized recommendations for each evaluator. These recommendations are tailored to respond dynamically to each user and differ in real time based on the user's activities. Figure 4-3 depicts different aspects considered during the process of computational (described in Section 4.4) when an evaluator has developed a new adaptive system, the process of adding the new system, and the steps involved.



**Figure 4-3: Different aspects considered during the process of recommendation**

The algorithm takes into consideration four main factors during computations (presented in Table 4-1). For example, when recommending an evaluation approach to be used, the ingredients and factors considered include: (i) number of publications, (ii) type or venue of publication (journal, conference or workshop), (iii) the adaptive systems variation type, purpose or goal of evaluation and (iv) the number of selected evaluation purposes that are associated with the evaluation approach. The factors considered when computing which evaluation approach and technique to recommend are discussed in Section 4.4.1.

**Table 4-1: Ingredients used by the recommendation algorithm for an evaluation technique (methods, criteria and metrics)**

<b>Ingredients of the recommendation considered for evaluation techniques</b>	
	<b>1<sup>st</sup> Factors</b>
$N_p$ :	Number of publications published that used the evaluation method.
$T_p$ :	Total number of publications in the database
	<b>2<sup>nd</sup> Factors</b>
$T_j$ :	Total number of journal papers in the publications table
$T_c$ :	Total number of conference papers in the publication table
$T_w$ :	Total number of workshop papers in the publication table
$T_v$ :	Total venue score of all publications (i.e. calculated as: $4 * T_j + 2 * T_c + 1 * T_w$ )
$N_j$ :	Number of journal papers in which the evaluation method was used
$N_c$ :	Number of conference papers in which the evaluation method was used
$N_w$ :	Number of workshop papers in which the evaluation method was used
$N_v$ :	Venue score of the evaluation method (calculated as $4 * N_j + 2 * N_c + 1 * N_w$ )
	<b>3<sup>rd</sup> Factors</b>
$T_{SV}$ :	Total number of systems that belong to a given variation type (v)
$N_{SV}$ :	Number of systems belonging to the given variation type (v) that were evaluated using the evaluation method.
	<b>4<sup>th</sup> Factors</b>
$T_{EVP}$ :	Total number of evaluation purposes (goal) selected by the user (i.e. how many checkboxes the user checked on the screen)
$N_{EVP}$ :	Number of selected evaluation purposes that are associated with the evaluation method

Section 4.4.1 discusses the implementation of these algorithms and how these factors are computed to produce a recommendation. When computing the recommendations for evaluation techniques (i.e. methods), the algorithm uses the ingredients and factors shown in Table 4-1.

- ***Explanation Algorithm***

The explanation algorithm capabilities enable provision of explanations as to why each factor was taken into consideration and why that evaluation approach and technique was recommended. Explanations provide transparency, validity, trustworthiness, persuasiveness, effectiveness, efficiency, satisfaction, relevance, comprehensibility and

education (Tintarev and Masthoff, 2009). The recommended methods are then computed into a score. These scores are presented to users as stars; for example, the most appropriate method is ranked at the top with five stars and the least appropriate with one star.

- ***Knowledge-base***

Knowledge-base automated repositories are able to reason about how a particular item meets a particular user need. The candidate define a knowledge-base as an information repository that provides a means for information to be collected, organized, shared, searched and used. The knowledge-base contains internal models (published studies, system model, learning portal and user model) which are used to store data on existing evaluation cases. The dataset discussed in Chapter 5, section 5.3 is based on peer-reviewed evaluation cases. Thus, rather than being a large dataset based on many users' behaviour, it is based on a smaller dataset that has been quality-reviewed. Moreover, the dataset can grow over time as the framework itself provides a mechanism for published authors to add their evaluation cases to the dataset; thus the candidate believes this dataset is a very valuable dataset for AEL evaluation choices but will become even more so in the future.

- ***Recommended Evaluation Approach and Techniques***

The key ingredients used when recommending an evaluation approach are shown in Table 4.2. These include: (i) number of publications, (ii) type or venue of publication (journal, conference or workshop), (iii) the adaptive systems variation type, purpose or goal of evaluation, and (iv) the number of selected evaluation purposes that are associated with the evaluation method. The factors considered when computing which evaluation approach and technique to recommend are discussed in Section 4.4.1.

**Table 4-2: Ingredients used by the recommendation algorithm for evaluation approach**

<b>Ingredients of the recommendation considered for evaluation approach used</b>	
P:	Publication (published study)
S:	System
V:	Variation type (category)
E <sub>VP</sub> :	Evaluation purpose (goal)
N:	Number of ...
T:	Total number of...
<b>1<sup>st</sup> Factors</b>	
N <sub>P</sub> :	Number of publications published that used the evaluation method
T <sub>P</sub> :	Total number of publications in the database
<b>2<sup>nd</sup> Factors</b>	
T <sub>J</sub> :	Total number of journal papers in the publications table
T <sub>C</sub> :	Total number of conference papers in the publication table
T <sub>W</sub> :	Total number of workshop papers in the publication table
T <sub>V</sub> :	Total venue score of all publications (i.e. calculated as: $4 * T_J + 2 * T_C + 1 * T_W$ ).
N <sub>J</sub> :	Number of journal papers in which the evaluation method was used
N <sub>C</sub> :	Number of conference papers in which the evaluation method was used
N <sub>W</sub> :	Number of workshop papers in which the evaluation method was used
N <sub>V</sub> :	Venue score of the evaluation method (calculated as $4 * N_J + 2 * N_C + 1 * N_W$ )
<b>3<sup>rd</sup> Factors</b>	
T <sub>SV</sub> :	Total number of systems that belong to a given variation type (v)
N <sub>SV</sub> :	Number of systems belonging to the given variation type (v) that were evaluated using the evaluation method
<b>4<sup>th</sup> Factors</b>	
T <sub>EVP</sub> :	Total number of evaluation purposes (goals) selected by the user (i.e. how many checkboxes the user checked on the screen)
N <sub>EVP</sub> :	Number of selected evaluation purposes that are associated with the evaluation method



## User Input / Output

The architecture can not only recommend the approach but also provide explanations as to why that approach was recommended. Figure 4-4 depicts different aspects considered (step 1 to 7) before recommending the top five most appropriate approaches. The first is the most appropriate one and the fifth the least.

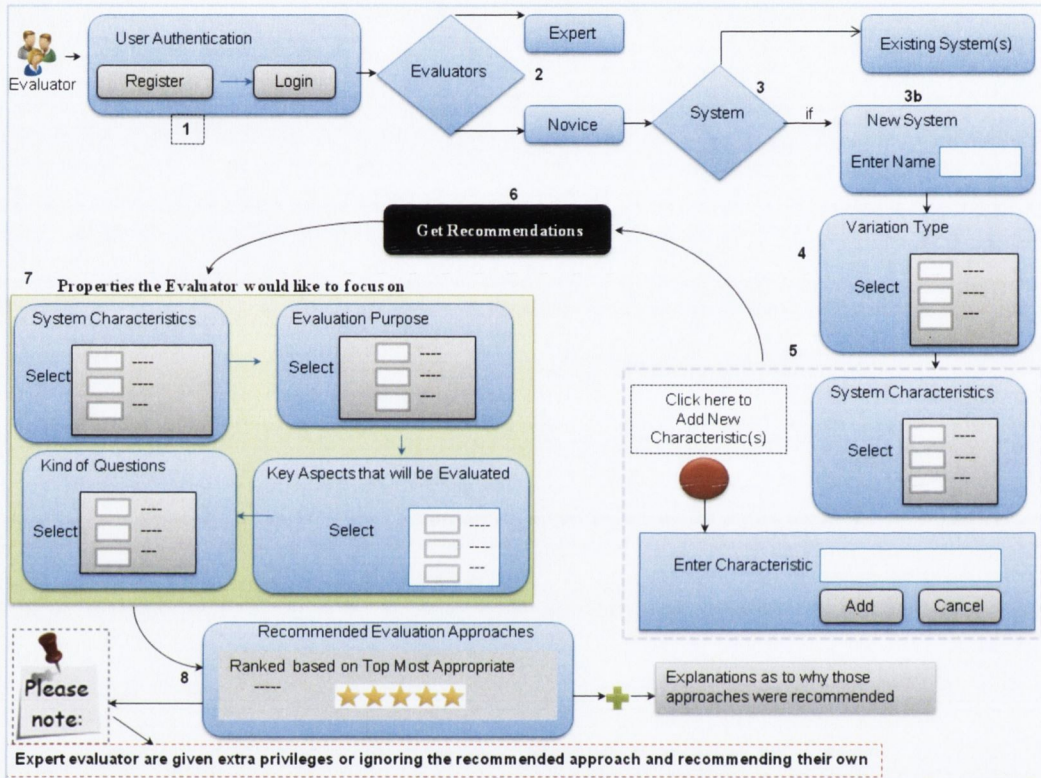
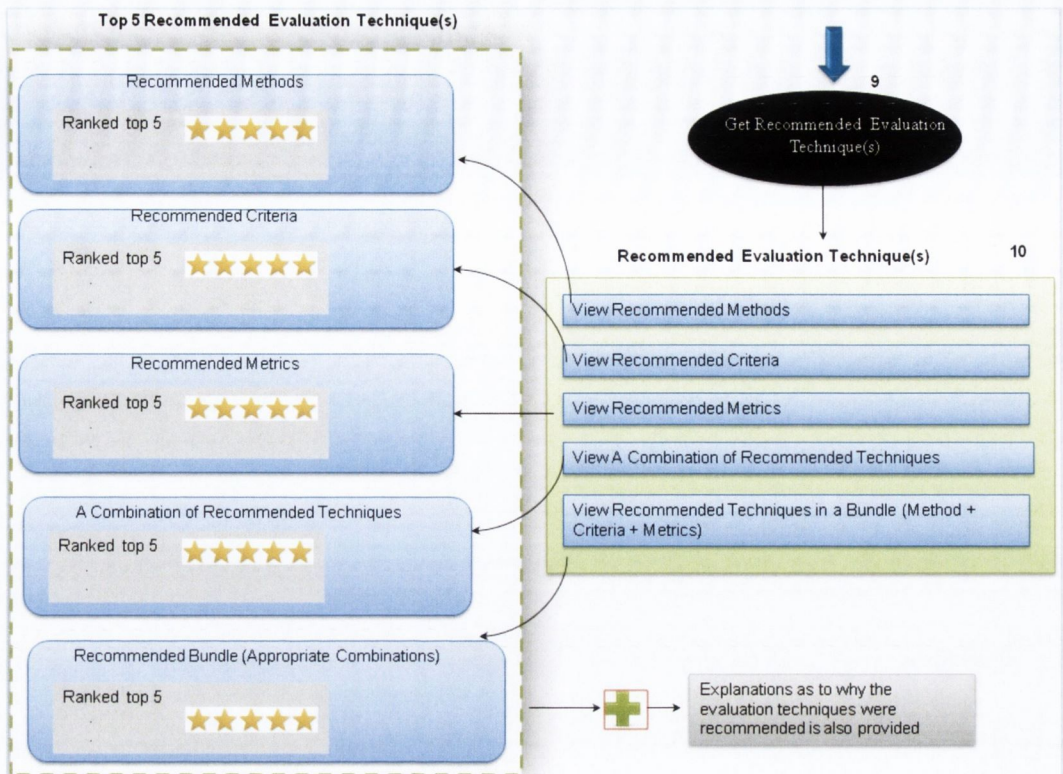


Figure 4-4: Different aspects considered during Process of recommending an approach

Once the user has received the recommended approach, the architecture can recommend evaluation techniques (method, metric and criteria) once the user triggers GetRecommendedTechniques (depicted in Figure 4-5).



**Figure 4-5: User triggers 'GetRecommended evaluation technique(s)**

Figure 4-6 shows the process involved before an evaluation technique is produced:

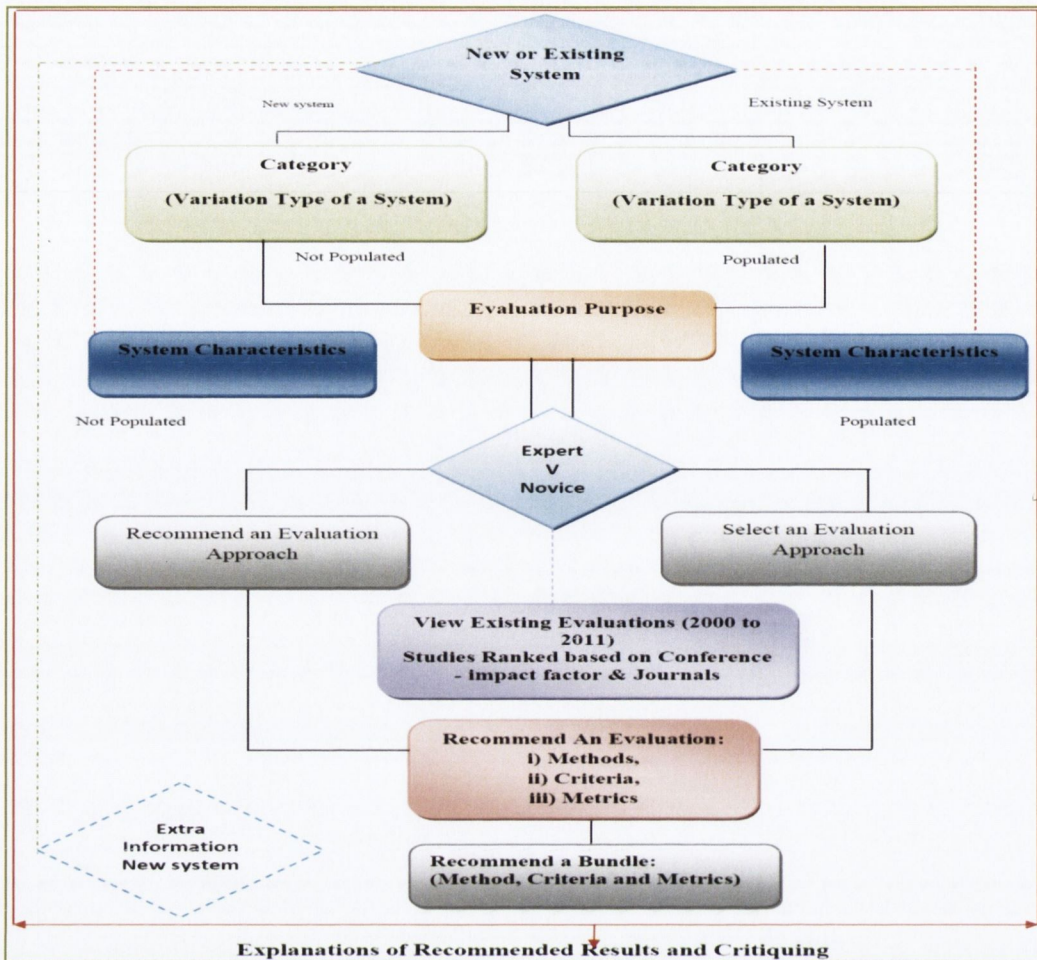


Figure 4-6: Process before a recommendation is produced

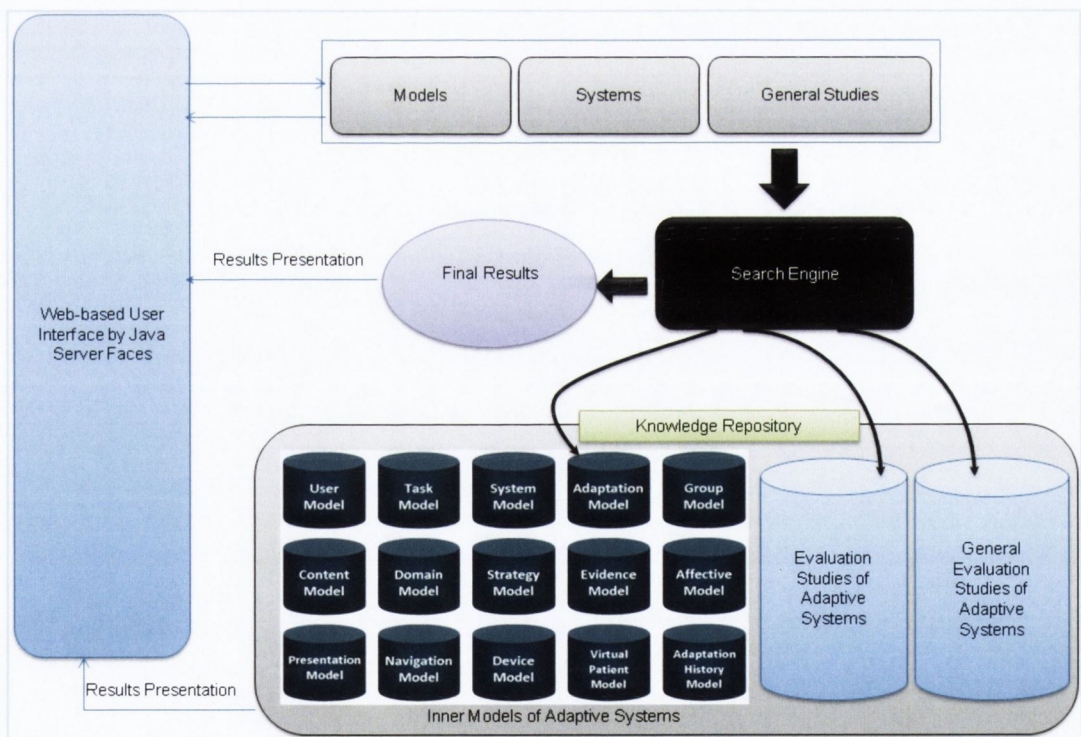
#### 4.3.1.2 Personalized Search Sub-System for Evaluation Studies of Adaptive Systems

Personalization can be based on many different attributes. These include: (i) user age, user disability, subject prerequisites, user role, user motivation, user language, preferred modality (speech, video), user prior knowledge, user competencies, user experiences/history, user objectives, user emotion, user preferences, user interests, user goals and user behaviour (Wade, 2009); (ii) the quality approach, the lifecycle approach, and (iii) display device relationship to other objects, time, performance, level of control, activity, process rules, interaction (with disciplines, group membership, group activity, deadline). One of the core issues in learning is the personalization of the learning experience. It is widely recognized that effective and efficient learning need to be individualized, personalized and learner-controlled. The dimensions of personalization

include: content models, subject domain models, environment models, service models, user and cultural models, activity models, implicit and explicit model triggers, artificial intelligence and non-artificial intelligence approaches. Furthermore multi-dimensional enables: personalized (collaborative) tasks, personalized situational simulations, personalized games, personalized mobile collaboration, personalized social networking and personalized community support.

Personalization for the end user improves user efficiency, user effectiveness and user satisfaction. It is important to review aspects of the personalization that need to be evaluated (e.g. quality of the user modelling, performance of different adaptation approaches, knowledge gain from using the personalized system, overall end-user experience). Several evaluation techniques need to be combined and executed differently.

This section discusses the proposed architectural design (Figure 4-7) capabilities of the personalized search sub-system of which the database is populated using 250 evaluation studies published since 2000.



**Figure 4-7: Architecture of the personalized search sub-system**

## ***Search Engine***

The architecture of the search engine is determined by two main requirements: the effectiveness (quality of results) and the efficiency (response time and throughput). The key capabilities include: performance with efficient search and indexing capabilities; adaptability, which involves tuning of the system, and scalability enabled by the engine growing with data and evaluators over a period of time. The search engine is responsible for computing and enabling the indexing process.

## ***Knowledge Repository***

The knowledge repository capabilities include storage of a centralized database of evaluation studies on: i) internal models of adaptive systems, ii) adaptive systems and iii) general evaluation studies.

## ***User Interface***

In this case the users are provided with an automated personalized search interface consisting of three distinct user interface components, which allow users to find (i) evaluation studies of the internal models (discussed in Section 2.3.2), (ii) evaluation studies of adaptive systems and the evaluation techniques used during evaluations, and (iii) general evaluation studies of adaptive systems. This interface extracts data stored in a centralized database. The user interface also allows users to interact online and query the search systems.

## ***User Input/Output***

The architecture also provides a centralized database which is populated using over 250 evaluation studies published since 2000. The users are provided with an automated personalized search interface, which allows them to find evaluations of 15 internal models of adaptive systems, evaluation of 106 adaptive systems developed since 2000 and over 130 studies of such systems.

### 4.3.1.3 Taxonomy of Technical Terms

The architecture (Figure 4-8) of the taxonomy has capabilities such as provision of the science of categorization of things based on a predefined system and contains a controlled vocabulary with a hierarchical tree-like structure (Liu F. et al., 2011). The proposed taxonomy consists of technical terms (identified during recommendation of evaluation approaches and techniques, discussed in Section 4.3.1.1) of evaluations of adaptive systems in general. It is both a hierarchical classification scheme and a controlled vocabulary of terms (alphabetically presented from A to Z) that identifies the content presented by the recommender system and personalized search systems.

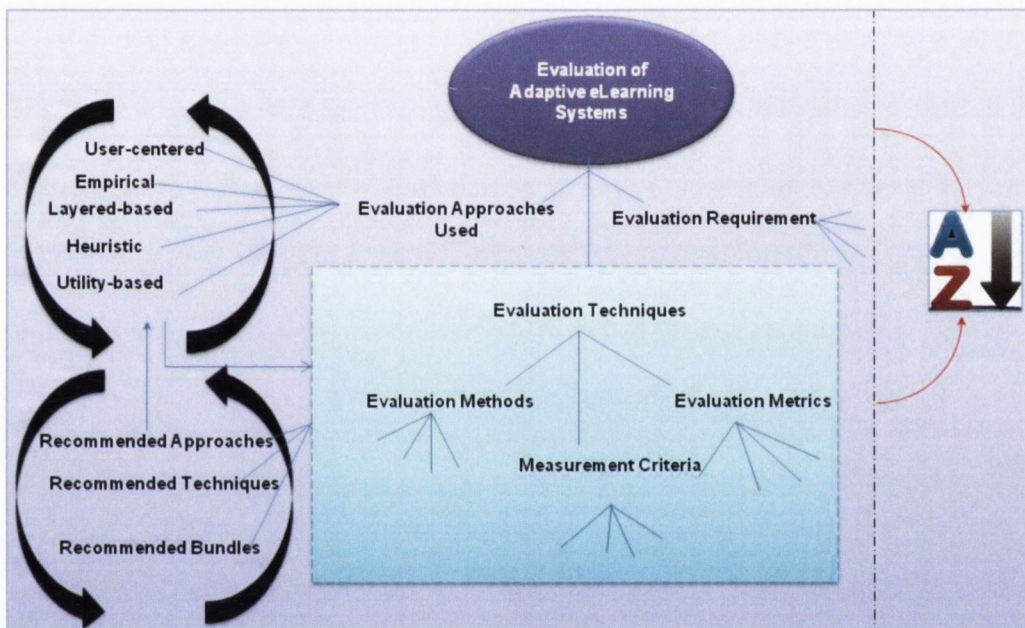


Figure 4- 8: Taxonomy of technical terms architecture

The overall EFEx technical design includes several well intergrated technologies and the functionalities of all the components discussed in section 4.3.1 (Appendix B5).

### 4.3.2 Design Testing

It is important to involve the users of a system in architectural design decisions which the developer makes before implementations start. After using the influences from the state of the art in designing EFEx components, a real-life user study was conducted aimed at identifying the usefulness of the framework and also what the community (user modelling,

adaptation and personalization,<sup>14</sup> adaptive hypermedia,<sup>15</sup> dataTEL recommender<sup>16</sup> and knowledge and data engineering group in Trinity School of Computing<sup>17</sup>) considered would help support novice evaluators of adaptive systems.

#### ***4.3.2.1 Experiment Objectives***

The aim of this experiment was to find out which of the features (i.e. recommender, search, taxonomy and explanation) of EFEx Framework, the participants would find/or consider useful. The candidate also wanted to involve users in deciding which design components of EFEx they considered more useful. This would give us a real picture of what they wanted.

#### ***4.3.2.2 Experiment Setup***

Emails were sent to all delegates of the UMAP<sup>18</sup>, recommender systems, hypertext conference, user modeling and AH mailing list, requesting them for their consent to take part in the evaluation experiment, along with a link to the online quantitative questionnaire. The survey consisted of 10 questions, in two sections. Questions 1 to 9 (an investigation of evaluations of adaptive systems developed from 2000 to 2011) and Question 10, introducing the components of the proposed web-based evaluation framework for end-user experience in adaptive systems (EFEx). The full experiment questions can be found in Appendix C

#### ***4.3.2.3 Results and Findings***

A total of 550 emails were sent out. A total of 96 people participated in the online survey. Another 14 participated in structured interviews in which the candidate used the same questionnaire. An overall total of 110 participated; these were researchers who were evaluators and developers of adaptive systems developed from 2000 to 2011. The results and findings of this study are presented in two sections. Questions 3 to 9 results are presented in Chapter 5, Section 5.2, which discusses evaluations of adaptive systems, and the Question 1, 2 and 10 results and findings are discussed in this section.

---

<sup>14</sup> <http://www.um.org/>

<sup>15</sup> <http://www.ht2011.org/>

<sup>16</sup> <http://recsys.acm.org/2011/>

<sup>17</sup> <http://kdeg.scss.tcd.ie/>

<sup>18</sup> Conference proceedings( 2001-2011)

## User Characteristics

The quantitative questionnaire revealed that 81% of the users had developed an adaptive system between 2000 and 2011 and were able to identify system-specific characteristics. The characteristics gathered from the questionnaire are shown in Table 4-3. As can be seen from this table, there are a number of differences among the various users.

**Table 4-3: Identification of user characteristics (e.g. expertise in adaptive systems development)**

Have you developed an adaptive system in the past (from 2000 to 2011)? (An adaptive system refers to a system which tailors its output, using implicit inferences based on interaction with the user)*		
	<i>Response per cent</i>	<i>Response count</i>
Yes	80.6%	50
No	19.4%	12
If you answered yes to this question, please provide:		
<b>i) Name of Adaptive System</b>		
<b>ii) Year the System was Developed, Other Details</b>		
Provided names of systems	62	
Skipped	34	

\* This question was aimed at identifying the user's domain expertise (e.g. developers)

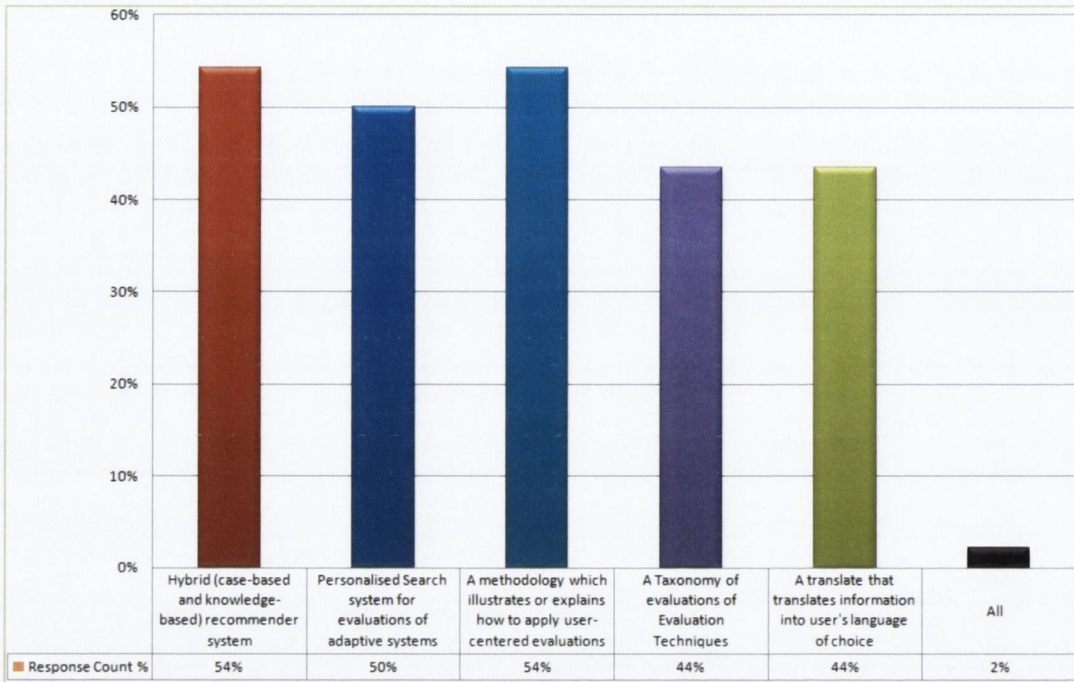
The results show the majority of the people sampled were actually developing adaptive systems. This is the exact community we wanted to survey.

The majority of the participants, 80.6%, indicated that they had developed an adaptive system. This was important because in Question 10 our target group of users were developers.

In Q10, the participants were presented with a list of five design components considered useful: (i) a repository for user-centred and layered evaluations of adaptive systems, (ii) recommendations on how to evaluate an existing adaptive system or a new adaptive system / internal models of adaptive systems, (iii) a user-centred evaluation methodology for adaptive systems, (iv) a taxonomy of evaluations of adaptive systems and (vi) information translated into user's language of choice.



They were then asked: *Which of the following features of EEx Framework would you find (consider) useful?* Figure 4-9 presents a summary of the results.



**Figure 4-9: Comparison rating of useful EEx components**

### *Usefulness<sup>19</sup> of EEx Framework Components*

When asked ‘*which of the following features of EEx Framework would you find (consider) useful?*’ 54.3% identified the most useful features to be: recommendations on how to evaluate an existing adaptive system or a new adaptive system, and inner models of adaptive systems, and a user-centred evaluation methodology for adaptive systems. This was followed by 50% selecting a centralized repository for user-centred and layered evaluations of adaptive systems and 43% taxonomy of evaluations of adaptive systems. Average was 17.5 (p=8.789). Only 2% indicated that they would find information translated into their language of choice useful. These findings enabled us to alter the overall design (Figure 3-4); the candidate decided to remove the translation component before starting implementations.

<sup>19</sup> Potential benefit and appropriateness of the features

Section 4.4 presents the implementation of the EEx architectural components considered useful, based on the outcomes of the survey.

## 4.4 Prototype Implementation

This section describes the implementation of the components of the architecture presented in Section 4.3. Section 4.4.1 describes the development of the hybrid recommender system component. During implementation of the user interfaces, JavaServer Pages (JSP) technology is used to create dynamically generated web pages based on HTML and XML. A use-case scenario of a novice evaluator getting recommendations on which evaluation approach to use and on evaluation techniques (method, metric and criteria) is also presented.

Section 4.4.2 describes the implementation of the personalized search system component. It also presents a use-case scenario of an evaluator searching for evaluation techniques for adaptive systems. Finally, the section describes the implementation of taxonomy of technical terms. The aim of this taxonomy is to support novice evaluators who encounter terminologies when using the hybrid recommender system. Figure 4-10 presents a screen shot of EEx evaluation framework user interface after the three components have been integrated together.

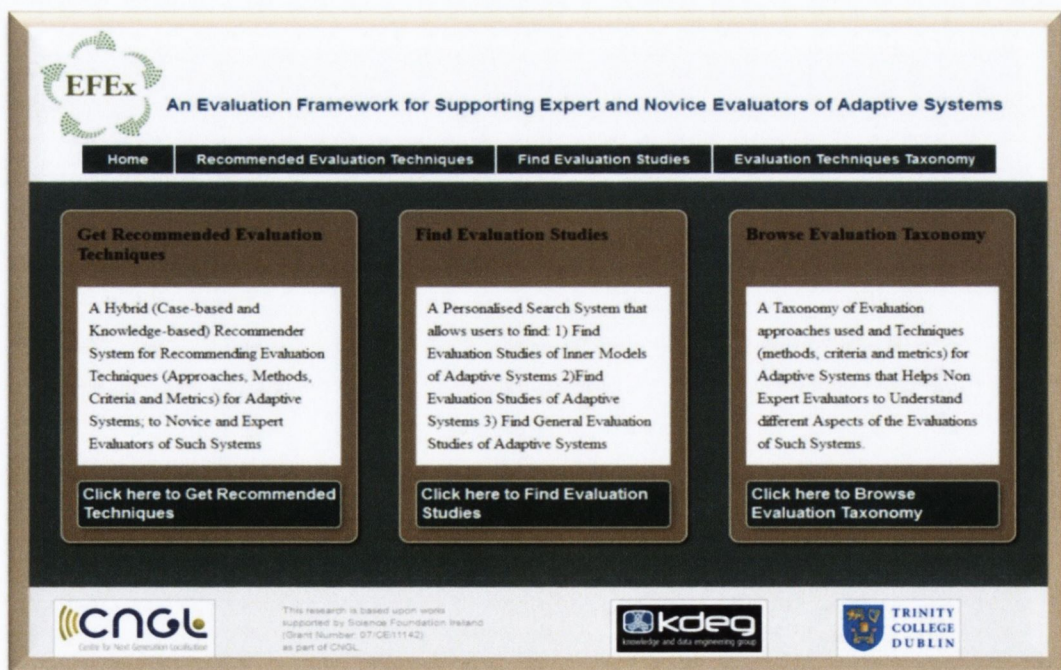


Figure 4-10: Automated EEx evaluation framework home page

This is the home page, where evaluators can decide which of the three components they prefer to visit first. When they select “Get Recommended Techniques”, they are automatically redirected to the hybrid recommender system main page; when selecting “Find Evaluation Studies”, they are redirected to the personalised search sub-system, while “Browse Evaluation Taxonomy” takes them to the taxonomy of technical terms (some of which they might have encountered when interacting with the recommender system). The following sections describe the implementation of these components.

#### **4.4.1 Hybrid (Case-based & Knowledge-based) Recommender**

##### **System**

The goal of developing any recommender component is to increase the effectiveness of problem-solving activity in scientific problem-solving environments. A hybrid (case-based and knowledge-based) recommendation technology reasoning techniques is used in order to enhance the appropriateness of suggestions of evaluation approaches and techniques for adaptive systems. More specifically the focus is on recommendation services for evaluating adaptive E-Learning systems. In particular, the multi-attribute relationships that people need to traverse when working out the most appropriate evaluation techniques are not easily navigated using typical database techniques.

A case-based reasoning (CBR) which is a computer technique that combines the knowledge-based support approach with a simulation of human reasoning, using past experience is used. The concept of case-based reasoning is founded on the idea of using explicit, documented experiences to solve new problems. The decision-maker uses previous explicit experiences, called ‘cases’, to help solve a present problem. The appropriate case is retrieved from the larger set of cases. The similarity between a present problem and the retrieved case is the basis for the latter’s selection (Mansar et al., 2003, Lorenzi12 et al., 2005). Figure 4-11 depicts the process involved in CBR; the knowledge cases are structured and stored in a case base, which the evaluator queries when trying to solve a problem.

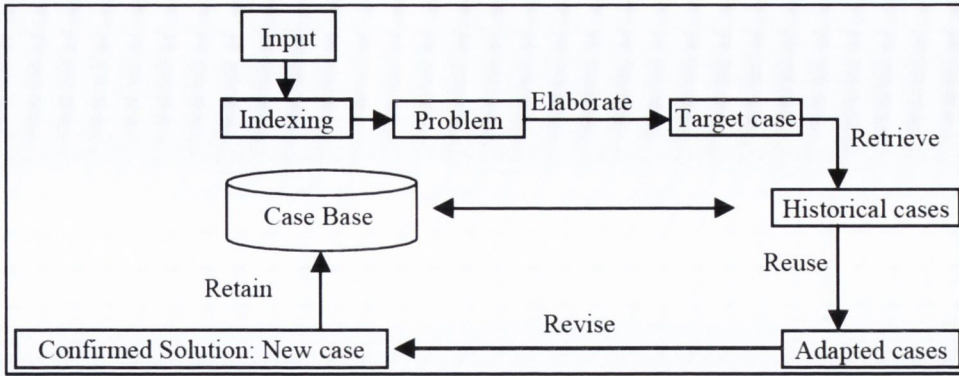


Figure 4-11: the CBR cycle, adapted from Choy et al., 2003

Items are retrieved using similarity measures (i.e. distance similarity):

$$\text{similarity}(p, REQ) = \frac{\sum_{r \in REQ} w_r * \text{sim}(p, r)}{\sum_{r \in REQ} w_r}$$

defined as:

- Sim (p, r) expresses for each item attribute value  $\phi_r(p)$  its distance to the customer requirement  $r \in REQ$ .
- $w_r$  is the importance weight for requirement r

Furthermore, the knowledge-based recommendation technique is used because it does not have a ramp-up problem since its recommendations do not depend on a base of user ratings. The ramp-up problem is a well-known issue with the content-based systems and recommendation systems in general. For example, new items cannot be recommended to any user until they get some sort of rating. Furthermore, it does not have to gather information about a particular user because its judgments are independent of individual tastes. These characteristics have been shown to make knowledge-based recommenders not only valuable systems on their own but also highly complementary to other types of recommender systems.

In this thesis, the candidate has developed a hybrid recommendation system that combines the two recommendation techniques to gain better system optimization, with fewer of the weaknesses of any individual ones. The implementation was divided into five main components: (i) user modelling, (ii) recommending evaluation approaches, (iii) recommending evaluation methods, criteria and metrics, (iv) recommending a combination

(bundle) of evaluation techniques, and (v) explanations as to why a particular approach or technique was recommended.

#### 4.4.1.1 User Modelling: Authentication

User modeling is a subdivision of human computer interaction; it describes the process of building up and modifying a user model. The main goal of user modeling is customization and adaptation of systems to the user's specific needs. A dynamic user model is used in order to collect and store personal data (i.e. organisation name, email address, username and password) associated with a specific evaluator.

For an evaluator to interact with the recommender system, they all have to register and then log in. Explanations are provided as to why authentication is required. This component is important; if the evaluators encounter any difficulties when using the system, then the candidate can trace the source of the problem. For example, if a user has registered and then forgotten their password, these details can be retrieved from the database. During the process of registration, evaluators are requested to provide organisation name, email address, username and password. Once registered, they are redirected to the login. Figure 4-12 presents a screenshot of the automated user authentication page.

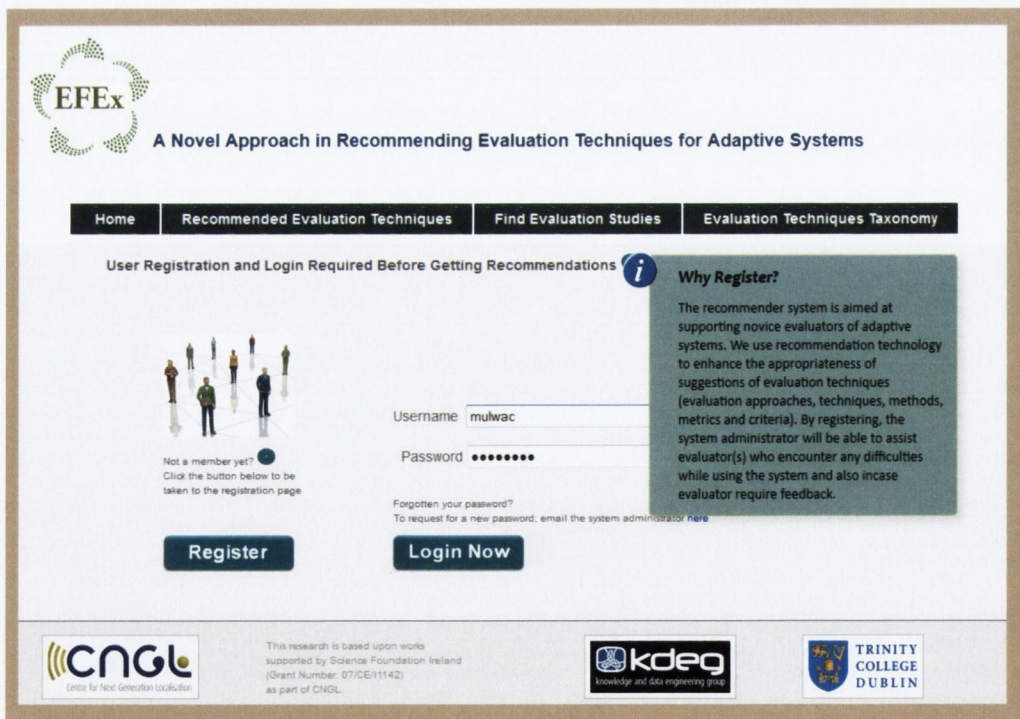


Figure 4-12: User authentication before interacting with the recommender system

#### 4.4.1.2 Evaluation Approaches Used

It is crucial that software developers and evaluators evade well-known pitfalls and that writers of future evaluation reports increase their empirical value by reporting the evaluation approaches used.

#### Expertise Identification Process

Before recommendations can be computed, the users are required to identify themselves: whether they are a novice (less than three years of experience or no experience in evaluations) or an expert evaluator (three years or more of experience in evaluations). If they are expert evaluators, more options are offered in order to customize recommendations. The developed user interface is shown in Figure 4-13, which presents a screenshot of users being asked whether they are a novice or an expert evaluator. Explanations are provided concerning why they are asked to identify themselves.

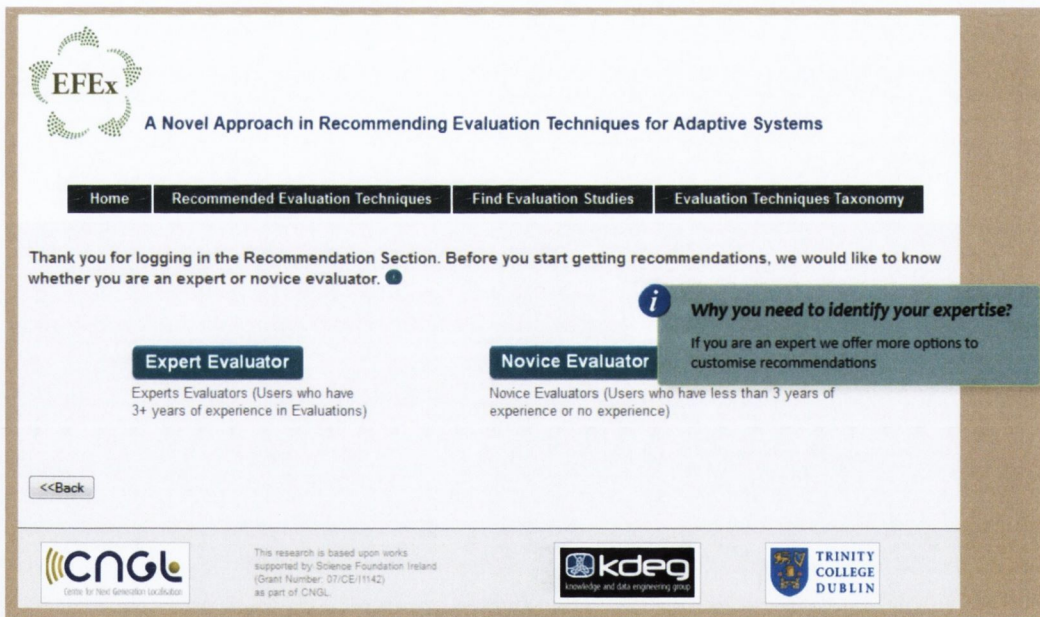
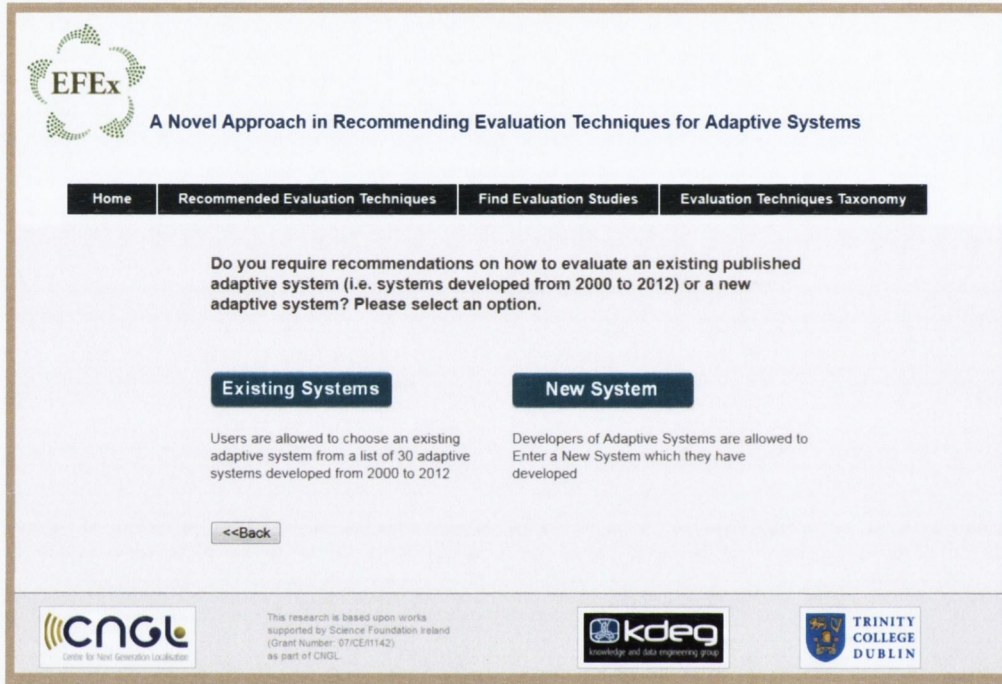


Figure 4-13: Expertise identification (novice or expert evaluator)

The next step of implementation is divided into two distinct features depending on whether the users identify themselves as a novice or an expert evaluator.

## Implementing Recommendations for a Novice

Before recommendations can start, the evaluator must state whether they require recommendations on how to evaluate an existing system (i.e. systems developed from 2000 to 2013) or a new adaptive system. They are presented with two options and requested to select one (Figure 4-14).



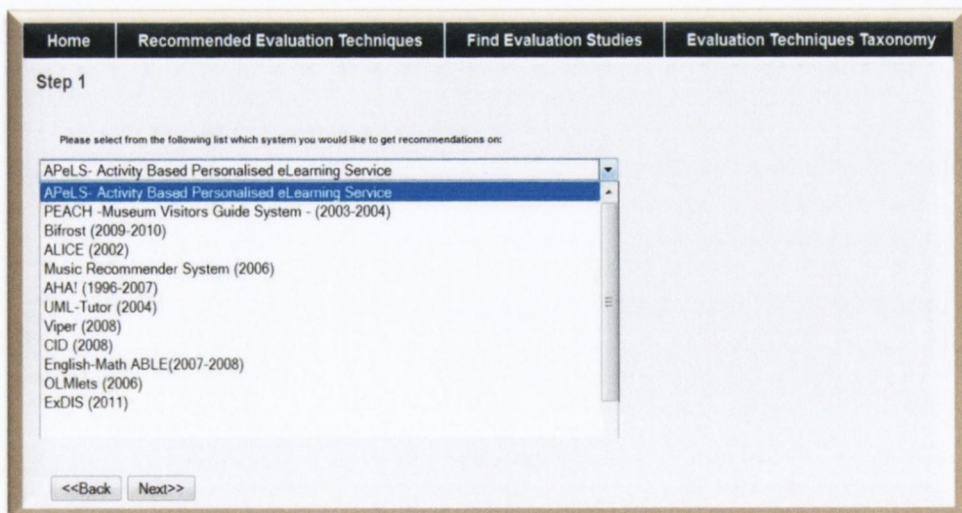
**Figure 4-14: Novice evaluator options: new system or existing system**

If the evaluator needs recommendations on existing systems, the recommender algorithm is applied to obtain the recommendations. There are six steps involved before a recommendation is produced. In steps 7 to 8 the user is presented with the recommended techniques. Table 4-4 presents a brief description of each step, assuming that a user has decided and selected an adaptive system (e.g. APeLS – Activity Based Personalized eLearning Service system) from a list of 30 existing adaptive systems, which are a subset of the 105 systems introduced in chapter 2 and chapter 7 sections 7.3.

**Table 4-4: Steps involved before recommending an approach**

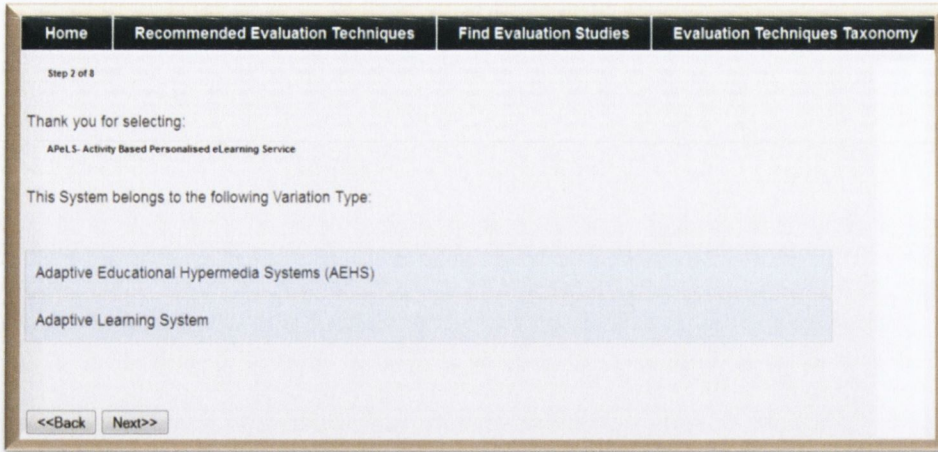
1 of 8:	Please select from the following list which system you would like to get recommendations on:
2 of 8:	Thank you for selecting: <b>APeLS – Activity Based Personalised eLearning Service</b> This System belongs to the following Variation Type:
3 of 8:	From this step to step 6, in order for us to provide you with recommendations, I would like to know which properties you would like recommendations on. Based on the characteristics of your adaptive system below, identify the ones you would like to focus on during recommendations. Please select one or more characteristics.
4 of 8	What are the goal(s) or purpose(s) of the evaluation that you are considering? Please select the evaluation purposes (more, many or all) from the list below:
5 of 8:	Based on your choices, listed below are key aspects which will be evaluated. Please choose the ones which you want to focus on evaluation.
6 of 8	Below are listed a set of questions which I consider to be useful for the evaluation that you are conducting. Click as many as required as you feel would be of interest to you:
7 of 8	Recommended approaches

Figures 4-15 to 4-20 present screenshots covering steps 1 to 6 before a recommendation is produced:

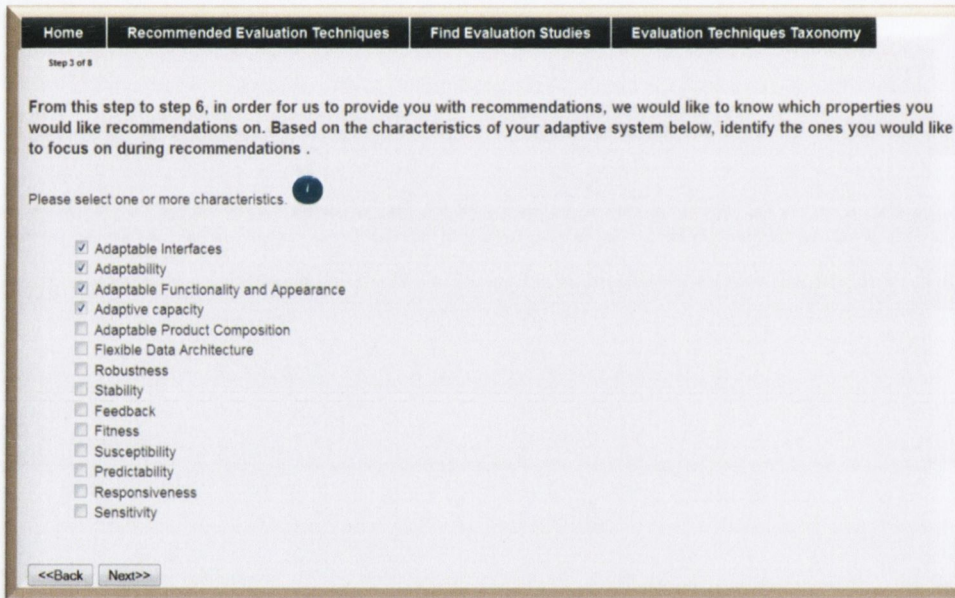


**Figure 4-15: Step 1 – System selection so that recommendations can start**





**Figure 4-16: Step 2 – Variation type of the system selected in step 1**



**Figure 4-17: Step 3 – Properties to select for focus during recommendations**

Home	Recommended Evaluation Techniques	Find Evaluation Studies	Evaluation Techniques Taxonomy
------	-----------------------------------	-------------------------	--------------------------------

Step 4 of 8

What are the goal(s) or purpose(s) of the evaluation that you are considering. Please select the evaluation purposes (more, many or all ) from the list below: ●

- Evaluate System Effectiveness
- Test End User Experience
- Test System Performance
- Test Adaptivity
- Test Usability
- Check Quality of Raw Input Data
- Check that Input Data is interpreted correctly
- Estimate Efficiency of the Instruction
- Check that Constructed Metadata Models Accurately Represent Real World
- Determine whether the Adaptation Decisions Made are the Optimal Ones
- Determine whether the Implementation of the Adaptation Decision made is Optimal
- Evaluate the Overall Adaptation Theory
- Summative Evaluation of the Adaptation Theory

<<Back   Next>>

**Figure 4-18: Step 4 – Selecting of evaluation purpose**

Home	Recommended Evaluation Techniques	Find Evaluation Studies	Evaluation Techniques Taxonomy
------	-----------------------------------	-------------------------	--------------------------------

Step 5 of 8

Based on your choices; listed below are key aspects which will be evaluated. Please choose the ones which you want to focus on evaluation ●

- Accuracy of recommendations
- System Efficiency
- System Effectiveness
- System Performance
- End user experience
- Accuracy of retrieved information
- Time taken to perform a task
- Learning experience
- Quality of raw input data

<<Back   Next>>

**Figure 4-19: Step 5 – Selection of key aspects to focus on, based on evaluator’s choices**

Home	Recommended Evaluation Techniques	Find Evaluation Studies	Evaluation Techniques Taxonomy
Step 6 of 8			
Below are listed a set of questions which I consider to be useful for the evaluation that you are conducting. Click as many as required as you feel would be of interest to you:			
<input checked="" type="checkbox"/> What is Improved by Having Adaptivity? <input checked="" type="checkbox"/> Why are you Testing for End-User Experience <input type="checkbox"/> Questions on Modelling the Current State of the World <input checked="" type="checkbox"/> How to Evaluate Adaptation as a Whole <input type="checkbox"/> Questions Relating to All Layers of the Adaptive System <input type="checkbox"/> Are adaptations done in a way that fits with user's expectations from the real world? <input checked="" type="checkbox"/> Can the user undo or change system interpretations, user modelling actions, adaptation decisions? <input type="checkbox"/> How appropriate was the action the system decided upon given the interaction state (and history) and the systems adaptive theory? <input type="checkbox"/> How necessary was the action the system decided upon? <input type="checkbox"/> Is the user informed about the kind of data the system captures about them, the type of inferences drawn and decisions <input type="checkbox"/> Does the system allow users to make unexpected pleasant discoveries, rather than restricting experience? <input checked="" type="checkbox"/> Is the timing of systems actions appropriately adapted to users' activities and context?			
<input type="button" value=" &lt;&lt;Back"/> <input type="button" value=" Next &gt;&gt;"/>			

**Figure 4-20: Step 6 – Selecting questions useful for the evaluation**

In step 7, the recommender algorithm and ranking algorithm are applied in order to obtain recommended approaches. To compute the recommendations, four cases (factors) were applied which were based on the candidates existing knowledge of evaluations of adaptive systems developed from 2000 to 2013:

- Number of publications in which the approach was used.
- Types of publications/venues (e.g. journal, conference, workshop) in which the approach had been used.
- How many adaptive systems belonging to the same variation type (category) had been evaluated using the same approach.
- Finally, the candidate gave extra weight to the approach according to its association with the selected evaluation purpose.

Table 4-5 presents a summary of the cases, values, weights and reasons considered when computing recommendations for a particular approach to be used.

**Table 4-5: Cases (factors), value, weight and reason considered when recommending an approach to be used**

Cases (Factor) Considered	Normalized Value	Weight	Explanation (Reason Narrative)
[1] Number of publications in which approach was used	$N_p / T_p$ <p>(i.e. N/Total number of publications)</p>	1	If N>0: “because this evaluation approach has been used. N out of T times. The approach has been used Np times out of Tp in the literature of evaluations of adaptive systems (2000-2011)”
[2] Types of publications/venues that the approach has been used in (e.g. journal, conference, workshop)	$N_v / T_v$ <p>i.e., {Journal = 4 ,Conference = 2 &amp; Workshop = 1 } Score</p> $\frac{\sum_{p=n}^{p=1} score}{\text{Total Venue score of all referenced publications}}$	1	“because the evaluation approach appeared in Nj Journals, Nc Conferences and Nw workshops”
[3] How many adaptive systems belonging to same variation type (category) have been evaluated using the same approach?  <i>*NB: where I have multiple variation types, I used the main variation type (not all of them)</i>	$N_{sv} / T_{sv}$ <p>(i.e., Get variation types, and then get No. of systems belonging to that variation type. No = of systems of the same variation type. Then get the No. of ones that used the approach)</p> $N_{sv} / T_v$ <p>(*NB: whether evaluated by this application or not)</p>	2	“because out of the Tsv systems which belonged to the “V” variation types (system categories) Nsv of them have been evaluated using the approach”

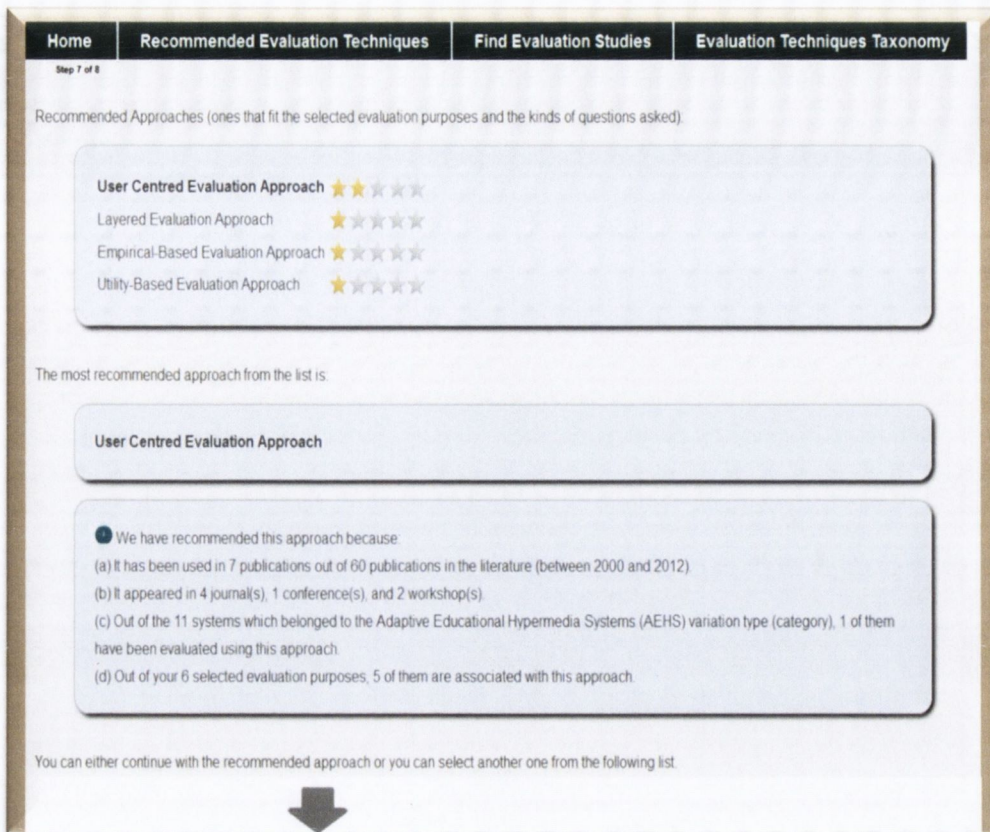
Table 4-6 Continued

Cases (Factor) Considered	Normalized Value	Weight	Explanation (Reason Narrative)
<p>[4] Give an extra weight to the approach according to its association with the selected evaluation purpose</p> <p><i>[Suppose I have multiple evaluation purposes, what will I do about it?]. How do I handle that? I account each evaluation type and increment.</i></p>	$N_{esv} / T_{evp}$ <p>(i.e., Count how many evaluation purposes that are associated with the approach were selected by the user).</p> $\frac{\text{No. of evaluation purpose selected } NP_{vp}}{T_{vp} \text{ total No. purposes selected by user or}}$	2	<p>“because, out of your Tep selected evaluation purposes, Nev<sub>p</sub> of them are associated with the evaluation approach”</p>

These cases (factors) were determined after an extensive review of evaluations of adaptive E-learning systems developed from 2000 to 2013(refer chapter 2). It was clear evidence that evaluation results of such systems were published in Journals, conferences and workshops.

To recommend the five most appropriate approaches, the candidate used a scoring method; the scores are out of four due to the cases (factors) weights presented in Table 4-5. To recommend the top most appropriate and accurate approaches, the score is then recalculated out five, so that the corresponding star rating (5 stars maximum) is displayed.

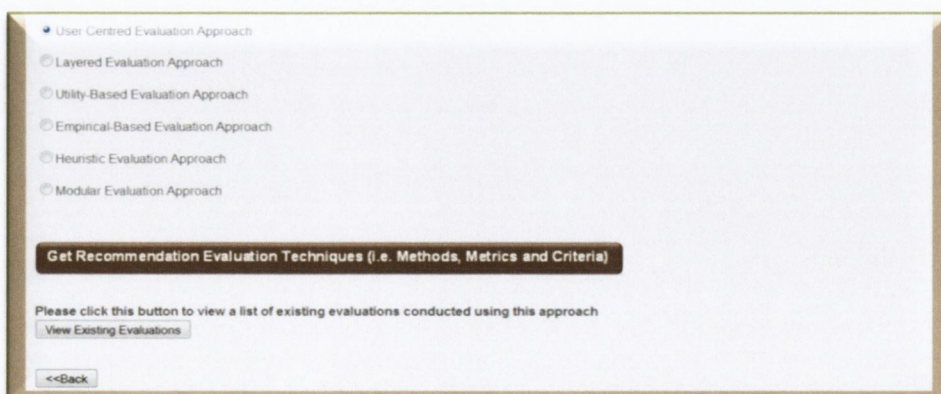
Finally, the evaluator is recommended five approaches that are appropriate to use when evaluating the system selected in step 1. The most appropriate approach is ranked at the top and highlighted in bold type, with the corresponding star rating. Figure 4-21 presents a screen shot of a recommended approach to be used when evaluating the **APeLS** System. The candidate then applied the explanation algorithm to explain to the evaluator why that approach was recommended.



**Figure 4-21: Top 5 recommended approaches and explanations**

### *Implementing Recommendations for an Expert Evaluator*

If an expert evaluator selects the same system (i.e. APeLS) and follows the same steps as the novice evaluator, he or she is offered more options to customise recommendations. Figure 4-22 presents the choices offered to the expert, with the recommended approach selected first.



**Figure 4-22: Expert evaluator offered more options to customize recommendations**

The next three sections describe the algorithms applied to obtain recommendations of evaluation techniques (methods, criteria and metric).

#### 4.4.1.3 Evaluation Methods

An evaluation technique can be considered as a combination or one or more evaluation methods and/or a set of criteria and/or one or more metrics. The recommender algorithm is applied to obtain a recommendation for techniques. For example an evaluation expert if he clicks “Get Recommended Evaluation Techniques”. He will be recommended a user-centered evaluation technique (shown in Figure 4-23).

The cases considered when computing the recommendations are presented in Table 4-6.

**Table 4-7: Factors considered when recommending an evaluation method**

Cases (Factor)	Normalized Value	Weight	Explanations
Number of publications in which the evaluation method was used	$Np/Tp$	1	“Because the evaluation method has been used $Np$ times out of $Tp$ times in the literature (2000-2012)”
The types of publications/venues that the evaluation method has been used in (e.g. journal, conference and workshop)	$Nv/Tv$	1	“Because the evaluation method appeared in $Nj$ journals, $Nc$ conferences & $Nw$ workshops”
How many adaptive systems belonging to the same variation type (category) have been evaluated using the evaluation method	$Nsv/Tsv$	2	“Because, out of the $Tsv$ systems which belonged to the “V” variation Type, $Nsv$ of them have been evaluated using the evaluation method”

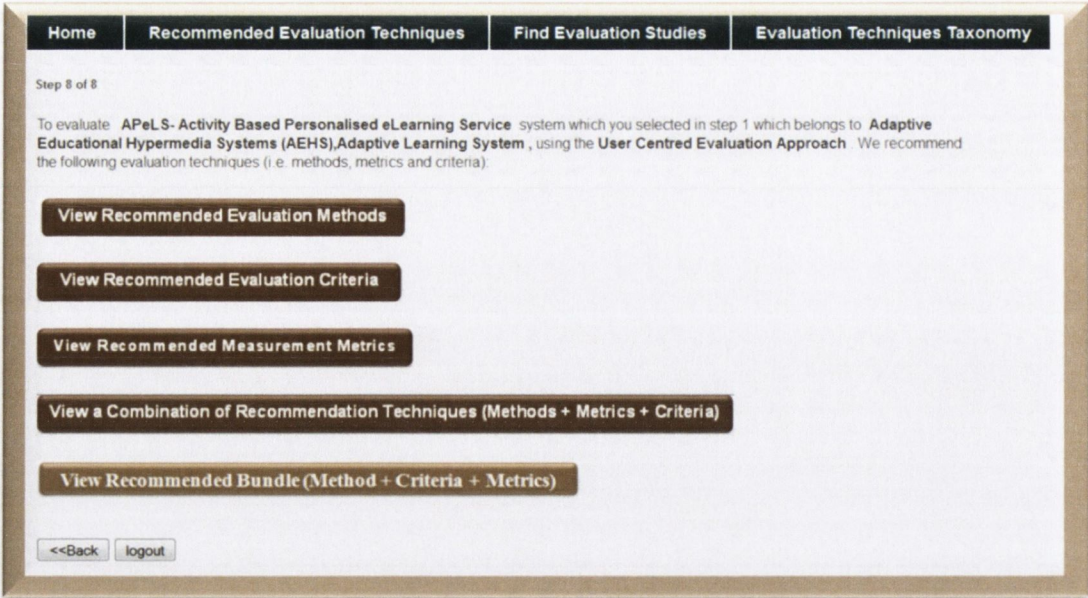
Table 4-7 presents the computations involved before a user can get recommended methods.

**Table 4-8: Computing recommendations for an evaluation method**

<b>Start</b>	Get all methods associated with the recommended approach in step 7 of 8.
Case (Factor) 1	<p>Number of publications in which the method was used: <math>N_p/T_p</math></p> <p>get <math>T_p</math> from database:</p> <p>get <math>N_p</math> for each method from database:</p>
Case (Factor) 2	<p>The types of publications (venues) that the method has been used in (i.e. journal paper, conference paper, workshop paper)</p> <p>get <math>T_j, T_c</math> and <math>T_w</math> from database:</p> <p>calculate <math>t_v</math>: <math>(int\ t_v = (4 * t_j) + (2 * t_c) + (t_w));</math></p> <p>get <math>N_j</math> for each approach from database:</p> <p>get <math>N_c</math> for each approach from database:</p> <p>get <math>N_w</math> for each approach from database:</p> <p>calculate <math>n_v</math> for each approach:</p>
Case (Factor) 4	<p>How many systems belonging to the same variation type have been evaluated with the method</p> <p>get first variation type id of the current system from the bean</p> <p>get <math>T_{sv}</math> from database:</p> <p>get <math>N_{sv}</math> for each approach from database:</p> <p><i>* Note: I only worked with the "first variation type for the system"</i></p>
Calculate score/sort	<p>Finally, calculate the total score for each method, then sort descending.</p> <p>1) The scores are out of 4 (due to the factor weights)</p> <p>Now I want to re-calculate score out of 5 instead, so that I can display the corresponding star rating (5 stars max)</p> <p>2) Approximate the score to for example 1.0 or 1.5 or 2.0 etc</p>
<b>End</b>	

Finally in step 8, the user can view the recommended methods by clicking “View Recommended Evaluation Methods” (Figure 4-23).





**Figure 4-23: View recommended evaluation methods, criteria and metrics, or a combination of the techniques and recommended bundle**

Figure 4-24 presents the ranked (top 5) recommended evaluation methods and explanations as to why that method was recommended.

Home Recommended Evaluation Techniques Find Evaluation Studies Evaluation Techniques Taxonomy

Step 8(b) of 8

**a) Recommended Evaluation Methods**

Post-test Questionnaire ★★★★★

The most recommended this method because  
 a) It has been used in 14 publications out of 60 publications in the literature (between 2000 and 2012).  
 b) It appeared in 6 journal(s), 5 conference(s), and 3 workshop(s).  
 c) Out of the 11 systems which belonged to the Adaptive Educational Hypermedia Systems (AEHS) variation type (category), 3 of them have been evaluated using this method.

Pre-and-Post-Test Questionnaire ★★★★★

The most recommended this method because  
 a) It has been used in 5 publications out of 60 publications in the literature (between 2000 and 2012).  
 b) It appeared in 3 journal(s), 1 conference(s), and 1 workshop(s).  
 c) Out of the 11 systems which belonged to the Adaptive Educational Hypermedia Systems (AEHS) variation type (category), 1 of them have been evaluated using this method.

Usability Testing ★★★★★

The most recommended this method because  
 a) It has been used in 17 publications out of 60 publications in the literature (between 2000 and 2012).  
 b) It appeared in 7 journal(s), 9 conference(s), and 1 workshop(s).  
 c) Out of the 11 systems which belonged to the Adaptive Educational Hypermedia Systems (AEHS) variation type (category), 5 of them have been evaluated using this method.

Interview ★★★★★

The most recommended this method because  
 a) It has been used in 17 publications out of 60 publications in the literature (between 2000 and 2012).  
 b) It appeared in 7 journal(s), 9 conference(s), and 1 workshop(s).  
 c) Out of the 11 systems which belonged to the Adaptive Educational Hypermedia Systems (AEHS) variation type (category), 5 of them have been evaluated using this method.

User Observation ★★★★★

The most recommended this method because  
 a) It has been used in 4 publications out of 60 publications in the literature (between 2000 and 2012).  
 b) It appeared in 2 journal(s), 2 conference(s), and 0 workshop(s).  
 c) Out of the 11 systems which belonged to the Adaptive Educational Hypermedia Systems (AEHS) variation type (category), 4 of them have been evaluated using this method.

**Please click this button to view a list of existing evaluations of systems belonging to the category(s) Adaptive Educational Hypermedia Systems (AEHS) Adaptive Learning System which have been evaluated using the same characteristics.**

View Existing Evaluations

<<Back logout

 This research is based upon works supported by Science Foundation Ireland (Grant Number: 07/CE11142) as part of CNGL  

**Figure 4-24: Ranked (Top 5) recommended evaluation methods**

If the user wants to view existing evaluations of systems belonging to the same category as the system he or she has been getting recommendations on, they can click the “View Existing Evaluations” button. A list of existing evaluation AEHS studies published from 2000 to 2013 is then presented to them.

#### 4.4.1.4 Measurement Criteria

The factors considered by the recommender and ranking algorithms for recommending the appropriate and accurate evaluation criteria are shown in Table 4-8.

**Table 4- 9: Factors considered when recommending criteria**

Factor	Normalized Value	Weight	Explanations
Number of publications/venues in which measurement criteria were used	$Np/Tp$	1	“Because the measurement criteria has been used $Np$ times out of $Tp$ times in the literature (2000-2012)”
The types of (venues) publications that the measurement criteria has been used in (e.g. journal, conference and workshop)	$Nv/Tv$	1	“Because the measurement criteria appeared in $Nj$ journals, $Nc$ conferences & $Nw$ workshops”
How many adaptive systems belonging to the same variation type (category) have been evaluated using the measurement criteria	$Nsv/Tsv$	2	“Because, out of the $Tsv$ systems which belonged to the “V” variation Type, $Nsv$ of them have been evaluated using the measurement criteria”

The same steps (as in Table 4-7) are applied but there is a difference with the cases:

- Case 1: Number of publications in which the criteria were used:  $Np/Tp$ .
- Case 2: Types of publications (venues) that the criteria have been used in (i.e. journal paper, conference paper, workshop paper).
- Case 3: How many systems belonging to the same variation type have been evaluated with the criteria,
- Finally, calculate the total score for each criteria and then sort it in descending order.

#### 4.4.1.5 Evaluation Metrics

The candidate applied the recommender algorithm to recommend the measurement metric. For example, in the final stage, to calculate the total score for each metric and then sort in descending order the calculations produced are depicted in JSP script (Appendix B4)

The steps are the same as the ones used when recommending an evaluation method (refer to Table 4-7) but there is a difference in the cases considered:

- Case 1: Number of publications in which the metric was used:  $N_p/T_p$
- Case 2: The types of publications (venues) that the metric has been used in (i.e. journal paper, conference paper, workshop paper)
- Case 3: How many systems belonging to the same variation type have been evaluated with the metric
- Finally, calculate the total score for each metric and then sort it in descending order.

When the evaluator clicks “View Recommended Metrics”, they are presented with the ranked (top 5) recommended metrics. Furthermore, explanations are provided as to why those metrics were recommended.

#### 4.4.1.6 Recommended Bundle (Method, Criteria & Metric)

For each recommended evaluation method (Section 4.3.1.1), a combination (bundle) of several evaluation criteria and metrics was created. The total numbers of bundles (i.e. method, criteria and metric) are ranked according to appropriateness. For each method the most appropriate evaluation criteria that occur most in the database with that method is selected. Then the most appropriate metric that occurs most with the method and the relevant criteria is also selected. The steps involved are:

<b>start</b>	
<i>Step 1</i>	<i>Connect to Database and prepare statements</i>
<i>Step 2</i>	<i>Execute a query to give the id of the method using the name of the method that the user chooses.</i>
<i>Step 3</i>	
3.1	<i>Execute the prepared query that retrieves the criteria id that has the maximum count for the method at hand from the previous-history database table</i>

3.2	Execute the prepared query that retrieves the metric id that has the maximum count for the method and criteria obtained from previous step.
3.3	Now, the user has the bundle elements: <i>method_id</i> , <i>criteria_id</i> , <i>metric_id</i> Execute three queries to get the names of the ids I have for the method, the criterion, and the metric.
3.4	Create a new <i>BundleItem</i> object with the three names (method name, criteria name, metric name).
3.5	Add the object of the <i>BundleItem</i> to the vector of bundles
Step 4	return the vector
<b>end</b>	

The factors considered are shown in Table 4-9.

**Table 4-10: Factors considered when recommending an evaluation metric**

Factor	Normalized Value	Weight	Explanations
Number of publications in which evaluation metric was used	$N_p/T_p$	1	“Because the evaluation metric has been used $N_p$ times out of $T_p$ times in the literature (2000-2012)”
The types of publications/venues that the evaluation metric has been used in (e.g. journal, conference and workshop)	$N_v/T_v$	1	“Because the evaluation metric appeared in $N_j$ journals, $N_c$ conferences & $N_w$ workshops”
How many adaptive systems belonging to the same variation type (category) have been evaluated using the evaluation metric	$N_{sv}/T_{sv}$	2	“Because, out of the $T_{sv}$ systems which belonged to the “V” variation Type, $N_{sv}$ of them have been evaluated using the evaluation metric”

Finally, in step 8 of 8, recommended bundles are produced when the user clicks “View Recommended Bundle” (Figure 4-25). A screen shot of recommended ranked (top 5) bundles when a user is getting recommendations for evaluating **APeLs** system.

Home Recommended Evaluation Techniques Find Evaluation Studies Evaluation Techniques Taxonomy

Step 8(e) of 8

Recommended Bundles (i.e. A Combination of Most Appropriate Evaluation Techniques (Evaluation Method + Criteria + Metric))

<p>Bundle#1 (Most Recommended)</p> <p>★ ★ ★ ★ ★</p>	<p><b>[ Post-test Questionnaire + Appropriateness of Adaptation + Precision ]</b></p> <p>This bundle is recommended because of the following. The method Post-test Questionnaire is the most recommended method. The criterion Appropriateness of Adaptation was the highest occurring criterion associated with the method in the reference database (used 1 times with the method). The metric Precision was the highest occurring metric associated with the method and the criterion in the reference database (used 1 times with the method and the criterion).</p>
<p>Bundle#2</p> <p>★ ★ ★ ★ ★</p>	<p><b>[ Pre-and-Post-Test Questionnaire ]</b></p> <p>This bundle is recommended because of the following. The method Pre-and-Post-Test Questionnaire is among the recommended methods. At the moment there are no criteria or metrics that are specifically recommended to be used with this method.</p>
<p>Bundle#3</p> <p>★ ★ ★ ★ ★</p>	<p><b>[ Usability Testing ]</b></p> <p>This bundle is recommended because of the following. The method Usability Testing is among the recommended methods. At the moment there are no criteria or metrics that are specifically recommended to be used with this method.</p>
<p>Bundle#4</p> <p>★ ★ ★ ★ ★</p>	<p><b>[ Interview + Perceived Usefulness + Accuracy of Retrieval ]</b></p> <p>This bundle is recommended because of the following. The method Interview is among the recommended methods. The criterion Perceived Usefulness was the highest occurring criterion associated with the method in the reference database (used 4 times with the method). The metric Accuracy of Retrieval was the highest occurring metric associated with the method and the criterion in the reference database (used 3 times with the method and the criterion).</p>
<p>Bundle#5</p> <p>★ ★ ★ ★ ★</p>	<p><b>[ User Observation ]</b></p> <p>This bundle is recommended because of the following. The method User Observation is among the recommended methods. At the moment there are no criteria or metrics that are specifically recommended to be used with this method.</p>

<<Back    logout

Figure 4-25: Recommended bundles (method, criteria and metric)

#### 4.4.1.7 Dataset for the Recommender System

The database is populated using an educational evaluation dataset that consists of a characterised, structured and interlinked list of evaluation studies of adaptive systems (60), adaptive systems (30), variation types (13), evaluation methods (74), evaluation criteria (75), and measurement metrics (85), and also evaluation approaches (7). The EFEx evaluation framework is built upon this dataset. The dataset is based on peer-reviewed evaluation cases. Thus, rather than being a large dataset based on many users' behavior. It is based on a smaller dataset that has been quality-reviewed. Moreover, the dataset can grow over time as the framework itself provides a mechanism for published authors to add their evaluation cases to the dataset. Thus the candidate believes that the dataset is already a very valuable dataset for adaptive E-Learning evaluation choices and will become more

valuable in the future. Running multiple recommender algorithms and systems over the dataset could provide a means of comparing recommender systems accuracy.

#### 4.4.2 Personalized Search System

The implementation of the personalised search system was divided into three components: (i) automated personalised search interface, (ii) search engine and (iii) knowledge repository.

##### *Automated Personalised Interface*

The importance of personalisation has been demonstrated by research in several areas, where human factors, such as level of knowledge, cognitive characteristics, purpose and goals have been shown to play an important role in providing successful personalisation (Magoulas et al., 2004). As discussed in Section 2.3.1, personalisation is very important, especially in improving the learning experience in TELE environments.

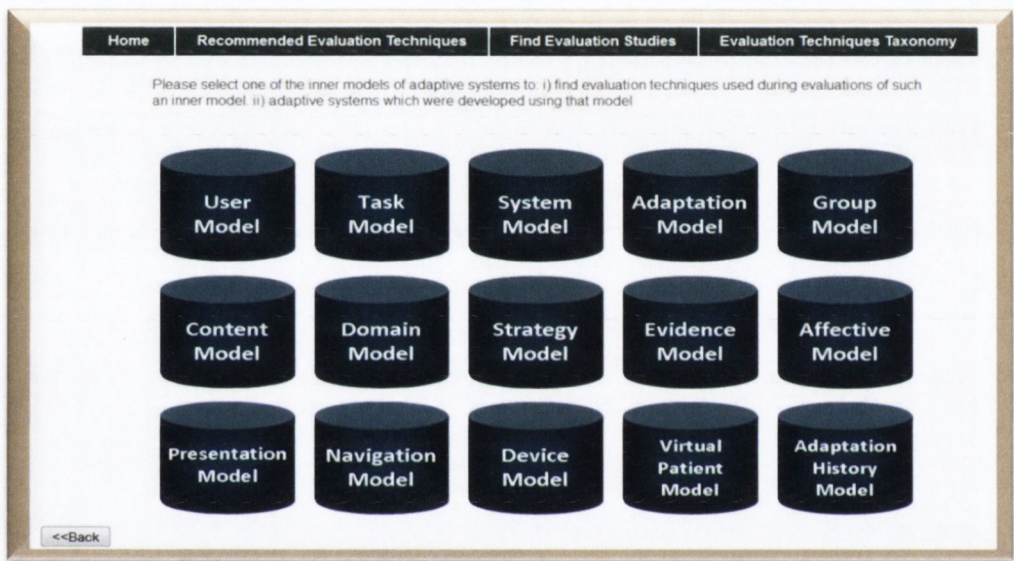
For this thesis the candidate developed an automated personalised search interface in order to support evaluators and help them find: (i) evaluation studies of internal models of adaptive systems, (ii) evaluation studies of adaptive systems and (iii) general evaluation studies of adaptive systems. Figure 4-26 shows these three interfaces.



Figure 4-26: Automated personalised search system UI

## Knowledge Repository

A knowledge repository was developed that systematically enabled us to capture, organize and categorize information that can be searched and data that can be quickly retrieved by evaluators interested in evaluation studies. For example, suppose a novice user is interested in finding evaluation studies of 15 different internal models of adaptive systems developed from 2000 to 2013: the user is directed to the user interface shown in Figure 4-27. Instead of all information about the internal models being presented at the same time, the evaluator sends a query search for the specific model they are interested in.



**Figure 4-27: Inner metadata models of adaptive systems**

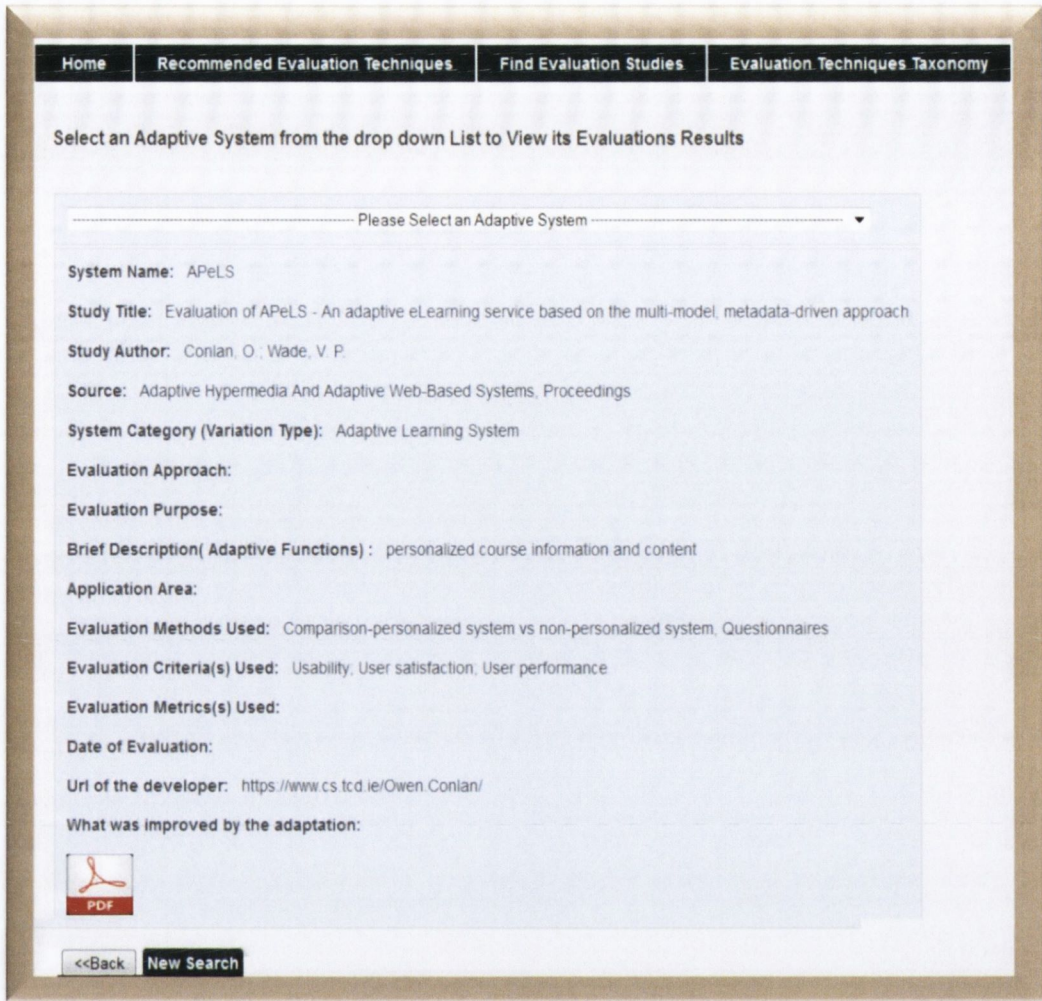
For example, suppose the evaluator clicked the model title "navigation model". The returned search results include the name of the model, evaluation methods, criteria and metrics used to evaluate that model and also a list of adaptive systems developed using that model. In this case the results returned are shown in Figure 4-28.



Home	Recommended Evaluation Techniques	Find Evaluation Studies	Evaluation Techniques Taxonomy
<b>Returned Search Results</b>			
Inner Model Name	Evaluation Method	Evaluation Criteria (Adaptive Factors Used)	Evaluation Metrics
Navigation	Comparison Questionnaires Interview,Heuristic	Perceived Usefulness, Usability, Appropriate of Adaptation Accuracy Performance	Precision, Accuracy of Models
Following is a List of Adaptive Systems Developed using such an the Inner Model			
CoMoLE ( 2007-2008)			
Dashboard at Kiwi Framework (2007-2010)			
Adaptive News System (2005)			
AdaptWeb (2001)			
NetCoach (2000)			
Activemath (2000)			
Inspire (2000)			
MyPlan (2008)			
Idiotypic control network for a navigating mobile robot (2006 ? 2009)			
<input type="button" value="←Back"/> <input type="button" value="New Search"/>			

**Figure 4-28: Returned query results for navigation model**

Suppose the evaluator is interested in finding evaluation studies of adaptive systems. The developed interface allows users to query the database (i.e. select from a drop list of 70 existing adaptive systems reported in the literature). If the evaluator selects the system called “APeLS”, the results returned are: system name, title of study, author, source, system variation type, brief description of the system, evaluation methods, criteria and metrics, and URL of the developer. Figure 4-29 presents the returned results of the users query.



**Figure 4-29: Results after query on APeLS system**

### 4.4.3 Taxonomy of Technical Terms

In addition the candidate created a taxonomy of technical terms used in the hybrid recommender system (Section 4.4.1) in order to help novice evaluators. It is an attempt to organize the various technical terms used during evaluations of adaptive systems and to relate them to the purpose and context of the systems. Users are provided with an automated user interface that consists of a glossary of terms classified in hierarchal and alphabetical views. The user interface is shown in Figure 4-30.

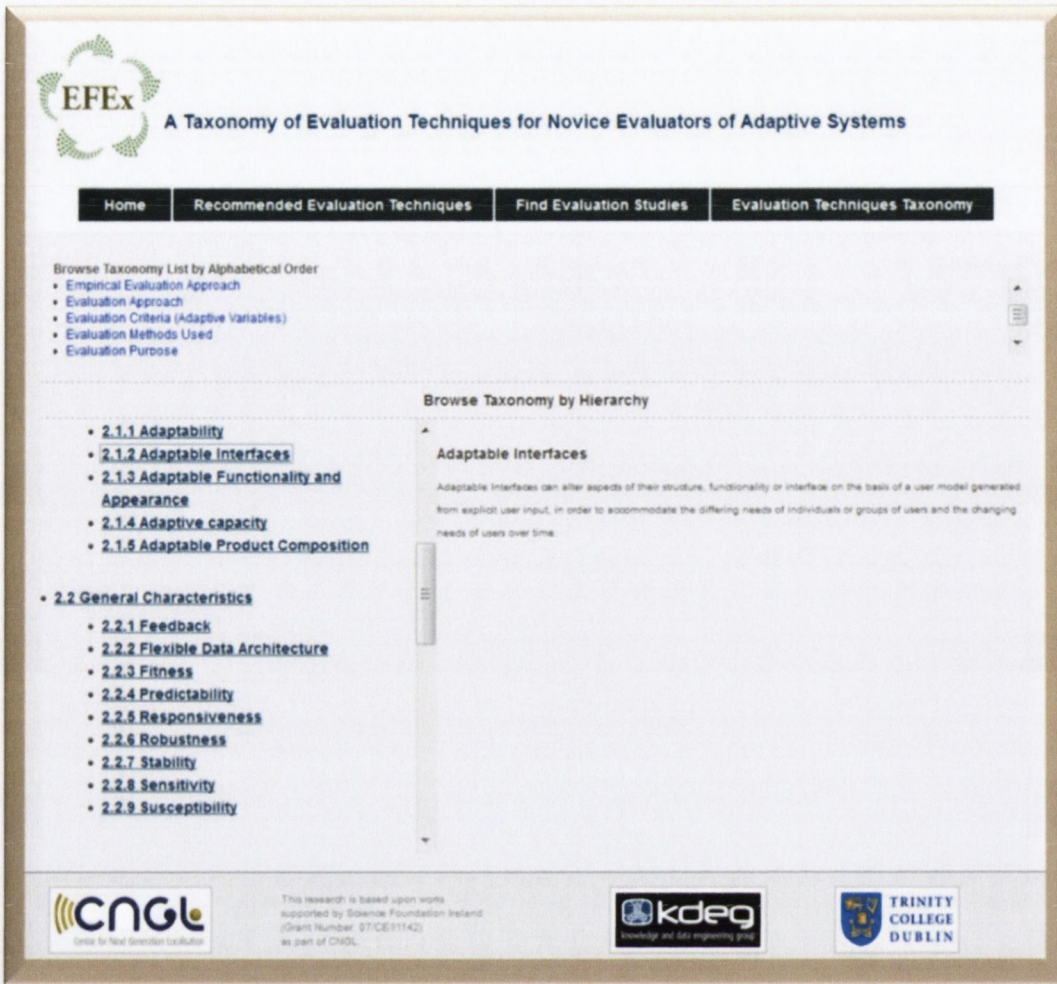


Figure 4-30: Taxonomy of technical terms

#### 4.4.4 Prototype Testing

It is important to evaluate any application before releasing it online to users. The first prototype of the EFEx framework was evaluated in terms of (i) perceived usefulness and (ii) usability from the user's perspective (user satisfaction).

##### 4.4.4.1 Evaluation Goal

##### *Perceived Usefulness and Usability*

The main goal of this experiment was to conduct a real-life study involving users from the modelling community in order to investigate the perceived usefulness of the EFEx prototype from the user's perspective.

#### 4.4.4.2 Evaluation Set-up

The experimental process (Figure 4-31) involved a demonstration of the prototype to users at the User Modeling, Adaption and Personalization, 19th International Conference (UMAP) 2011 (poster and demo section). A poster and demonstration paper of the EEx prototype accepted was accepted (Mulwa C. et.al., 2011).

#### Experimental Process

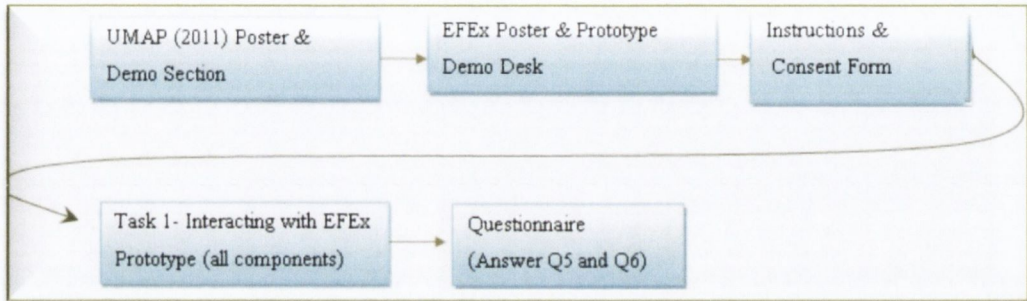


Figure 4- 31: Experimental Process

For every conference participant who stopped at our poster/demo desk, the candidate asked them whether they would consent to participate in a task that involved interacting with the EEx prototype. After using the framework, they were asked to complete a questionnaire which consisted of six questions. Table 4-10 presents the two questions about perceived usefulness and user satisfaction. The full questionnaire can be found in Appendix C.

Table 4-11: Task-based questions on usefulness of EEx

Q1. Having seen a demonstration of EEx, do you consider this framework is useful?					
Users	Very Valuable	Valuable	Somehow Valuable	At-least Valuable	Not Valuable
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Q2. Which of the following features of EEx did you find useful?					
a	Hybrid Recommendation System Component			<input type="checkbox"/>	
b	Personalised Search System Component			<input type="checkbox"/>	
c	UCE Methodology of adaptive systems			<input type="checkbox"/>	
d	A Taxonomy of Evaluation Techniques Component			<input type="checkbox"/>	
e	All			<input type="checkbox"/>	

#### 4.4.4.3 Results and Findings

A total of 25 users participated in this experiment. These were participants at the User Modeling, Adaption and Personalization, 19th International Conference (UMAP) 2011 (poster and demo section). The feedback from this group of participants was important because the community over the last 20 years has contributed immensely to the development and evaluation of adaptive systems research. This evaluation was part of the design and the candidate wanted to perform an earlier test to show usefulness of EFEx framework. However more depth of evaluation of EFEx usefulness is presented in chapter 6, section 6.2.1.3, section 6.3.3 and section 6.4.3.

After using EFEx, the participants were asked “Which of the following features of EFEx Framework would you find (consider) useful?” 85.2% agreed they found the framework useful. When asked to rate which features they considered valuable, using a rating scale of very valuable, somewhat valuable, not valuable, no response, they gave an average score of 1.52 for “very useful” and “useful” and an average score of 0.40 for “not useful”. These results are depicted in Figure 4-32.

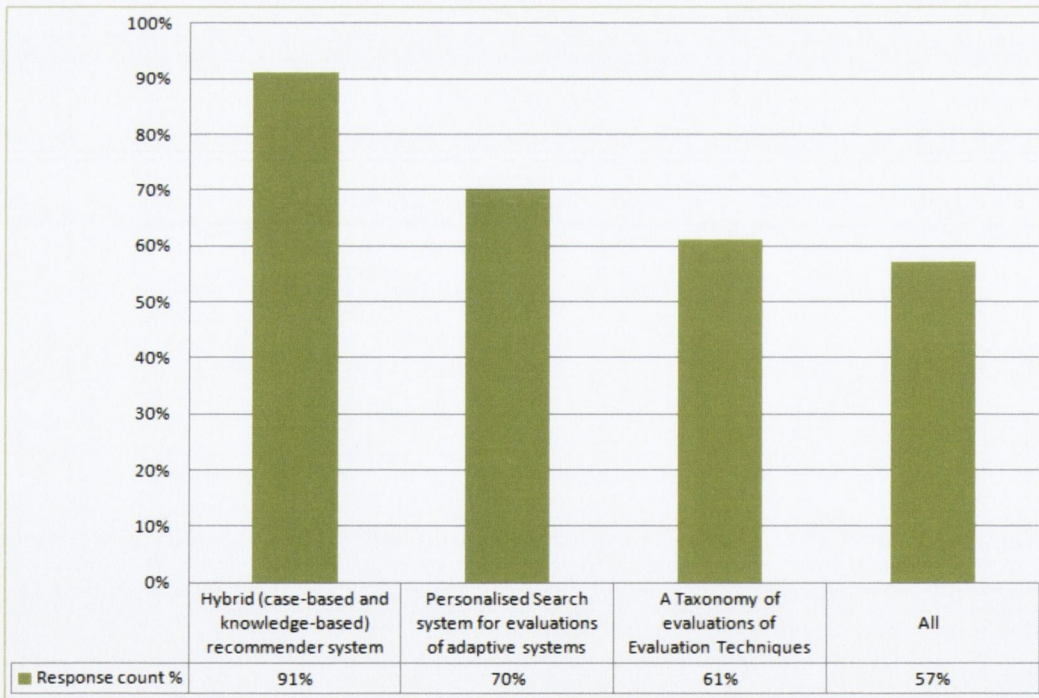


Figure 4- 32: Comparison of useful features of EFEx

The results confirm that the hybrid recommender system is the most useful feature of EFEx framework with overall response count of 91%, followed by the personalised search system and the taxonomy respectfully.

The candidate then computed the systems variation type results to find out the mean (average) and standard deviation, using the following formula:

Mean	Population Standard Deviation:	Variance (population standard deviation)
Mean = Sum of X values / N(Number of values)	$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$ <p> <math>\sigma</math> = population standard deviation  <math>x_i</math> = value of sample (i)  <math>\bar{x}</math> = mean of sample values  n = number of samples </p>	Variance = $s^2$

The population standard deviation for the dataset of variation types of adaptive systems was computed by

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - N)^2} = 6.16.$$

This shows that each data point in the sample sits an average distance of 6.16 statistical data points from the mean.

Where:

$\sigma$  = population standard deviation

$x_1 \dots, x_N$  = **variation** types of adaptive systems dataset

$\mu$  = mean of the variation types of adaptive systems population dataset

N = size of the variation types of adaptive systems population dataset

Dataset (approaches, methods, criteria and metrics) = (42, 32, 28, 26)

Total number = 4

The results after computing mean (average) and standard deviation:

Total number:	4
Mean (average):	32
Standard deviation:	7.12

This shows that each data point in the sample sits an average distance of 7.12 statistical data points from the Mean

## 4.5 Conclusion

By developing a fully automated web-based evaluation framework for supporting novice evaluators of adaptive systems (EFEx), the candidate has tackled research **Objective 3**. It is believed that the hybrid recommender system described in this chapter is novel and will be a very valuable support tool for novice evaluators of adaptive systems. Having conducted an evidence-based real life user study, which involved evaluators of adaptive systems developed from 2000 to 2013, it is clear there are no other similar hybrid recommendation systems in the field of adaptive E-Learning educational domain. This approach overcomes the limitations of case-based and knowledge-based approaches (discussed in chapter 2 section 2.3.3). The system is built on an educational evaluation dataset based on peer-reviewed evaluation cases (2000 to 2013), and can grow over time as the system provides a mechanism for published authors to add their evaluation cases to the dataset. The dataset can thus become progressively more valuable for evaluation choices in future. Furthermore, the personalised search system and the taxonomy of technical terms will provide extra support to evaluators of such systems.

In order to populate the database of the developed components of EFEx, the candidate conducted a real life user study survey which looked at “*how adaptive systems developed from 2000 to 2011 had been evaluated*”. The results and findings are discussed and presented in chapter 5.

# Chapter 5: Eliciting Knowledge-base for EFEx

## 5.1 Introduction

This chapter discusses the results of a real-life user study which was conducted to investigate evaluations of adaptive systems developed from 2000 to 2011. The chapter further presents the author's contribution of an education evaluation dataset for adaptive systems, created over a period of five years.

## 5.2 A User Study on Evaluations of Adaptive Systems Using an Evidence-based Approach

### 5.2.1 Evaluation Goal

The main goal of this evaluation was to investigate how adaptive systems have been evaluated over the past 10 years (2000 to 2011). Our aim was to tackle **Sub RQ1** – *‘What are the techniques used and tradeoffs between those techniques to support user-centered evaluations of adaptive systems?’* (Section 1.2)

By collecting evidence using a user study, from the adaptive systems scientific community on which systems were developed, what category they belonged to, and which evaluation techniques they had used during evaluations, the candidate would be able to get a clear picture on people's use of particular evaluation techniques and evaluation approaches.

### 5.2.2 Experiment Set-up

An experiment is a study in which at least one variable is manipulated and units are randomly assigned to the different levels or categories of the manipulated variable (Pedhazur and Schmelkin, 1991). The experiment setting involved sending emails plus a quantitative, structured online survey questionnaire<sup>20</sup> to five scientific communities (i.e.

---

<sup>20</sup> <http://www.surveymonkey.com/s/Q2DSDF8>



user modeling,<sup>21</sup> adaptive hypermedia,<sup>22</sup> recommender systems,<sup>23</sup> a knowledge and data engineering research group<sup>24</sup> and a centre for next generation and localization<sup>25</sup>). The term experiment setting is used in this research to mean a basic characteristic of evaluating research (Jannach et al., 2010).

A total of 500 emails were sent to the participants, with the URL link to the online questionnaire; 120 people responded; 96 of them participated online, while 24 answered the questionnaire in a lab set-up. Structured interview-based methods were used which involved asking the exact same questions we had asked to participants who participated online. The online questionnaire approach was suitable for investigating a wider range of overall evaluation approaches and techniques used in the evaluation of adaptive systems. Participants were required to complete nine closed questions. The full survey questionnaire can be found in Appendix C1.

### 5.2.3 Results and Findings

A total of 120 users participated in the study, out of which 110 completed the full evaluation by responding to all the questions. Participants were recruited from all the five scientific communities. Question 1 sought to gauge the interest of the user in adaptive systems. The aim of Q2 was to identify how many adaptive systems were developed from 2000 to 2011 and whether adaptivity had any impact on learners in adaptive TELE. Q3 aimed at identifying the facets of adaptivity and the impact of adaptivity on learners. Q4 and Q5 aimed at identifying which of the reported systems belonged to the TEL category and what metadata models were used. The most important questions were Q6, 7, 8 and 9, whose aim was to investigate how such systems were evaluated and the techniques (approaches, methods, metrics and criteria) used.

#### 5.2.3.1 Reported Adaptive Systems

To identify whether the participants had developed and evaluated an adaptive system, the following question was asked: “*Have you developed an adaptive system?*” A total of 80.6% responded yes (Figure 5-1).

---

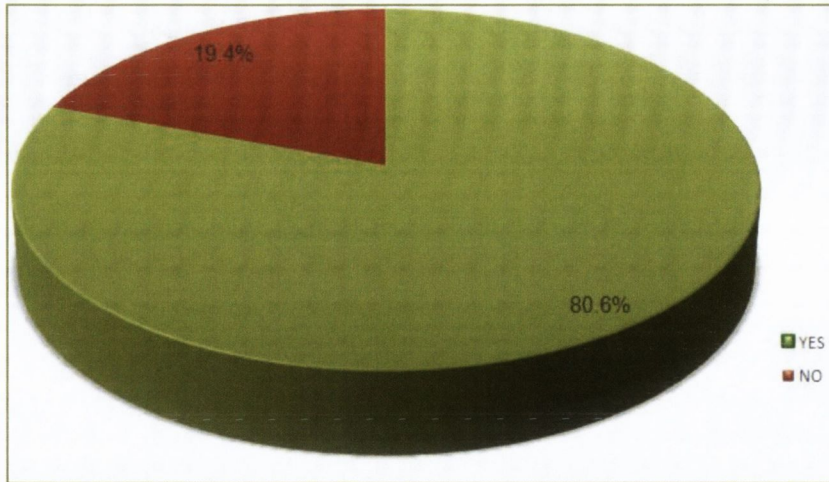
<sup>21</sup> <http://www.um.org/>

<sup>22</sup> <http://www.ht2011.org/>

<sup>23</sup> <http://recsys.acm.org/2011/>

<sup>24</sup> <http://kdeg.scss.tcd.ie/>

<sup>25</sup> <http://www.cngl.ie/>



**Figure 5- 1: Response to the question ‘Have you developed an adaptive system?’**

The next question was aimed at identifying the name of the system, the year the system was developed, what was improved by the system being adaptive, and the category each system belonged to (see Table 5-1). The aim of Q2 was also to identify how many adaptive systems had been developed from 2000 to 2011, so as to check whether adaptivity had any impact on learners after using the AEL system. Q3 aimed at identifying the facets of adaptivity and impact of adaptivity on learners. Q4 enabled us to identify which category the system belonged to.

**Table 5- 1: Response to question on name, year and category of reported system**

<b>Q2. If you answered yes to this question, please provide:</b>	
i) Name of Adaptive System,	
ii) Year the system was developed,	
iii) Other details	
Name of system	<b>70</b>
<b>Q3. If you have developed an adaptive system(s), What was improved by adaptivity?</b>	
Features improved by adaptivity	<b>50</b>
<b>Q4. What is the variation type of the adaptive system you have developed?</b>	
Adaptive Educational Hypermedia System	<b>69.4%</b>
Adaptive Information Retrieval System	38.9%
Online Help Customer Care System	11.1%

A total of 77 adaptive systems were reported. These systems belonged to several categories. Most were adaptive educational hypermedia systems. The response rate was 69.4%. The systems category results were computed to find out the mean (average) and standard deviation, using the following formula:

Mean	Population Standard Deviation	Variance (population standard deviation)
Mean = sum of X values / N (Number of values)	$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$ <p> <math>\sigma</math> = population standard deviation  <math>x_i</math> = value of sample (i)  <math>\bar{x}</math> = mean of sample values  <math>n</math> = number of samples                 </p>	Variance = $s^2$

The population standard deviation for the dataset of categories of adaptive systems was

computed by: 
$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - N)^2} = 7.70$$

Where:

$\sigma$  = population standard deviation

$x_1 \dots, x_N$  = categories of adaptive systems dataset

$\mu$  = mean of the categories of adaptive systems population dataset

$N$  = size of the categories of adaptive systems population dataset

Dataset = (25, 14, 4, 19)

Total number = 4

The results after computing mean (average) and standard deviation:

Total number:	4
Mean (average):	15.5
Standard deviation:	8.89

### 5.2.3.2 Reported Evaluations Techniques (Methods, Criteria & Metrics)

The methodologies for evaluating adaptive TEL systems are generally borrowed from the methodologies used in HCI and by those used for the evaluation of the information selection process (Gena, 2005).

#### **Evaluation Methods**

We asked participants if they had conducted a whole-system evaluation, and what evaluation methods they had used. A total of 35 methods were reported. Table 5-2 presents a summary of the reported methods.

**Table 5- 2: Evaluation methods reported in the study**

<b>Method</b>	<b>Percentage 75% to 21%</b>	<b>Method</b>	<b>Percentage 20% to 0%</b>
Questionnaires	75.0%	Prototyping	18.8%
Evaluation	60.4%	Heuristic Evaluations	14.6%
Interviews	50.0%	Creative Brainstorming Sessions	14.6%
User Observation	50.0%	Wizard of Oz Simulation	14.6%
Usability Testing	45.8%	Scenario-Based Design	8.3%
Data Mining	41.7%	Verbal Protocol	8.3%
User Test	35.4%	Grounded Theory	6.3%
Empirical Observations	35.4%	Discussion Group	6.3%
Quantitative	31.3%	Systematic Observation	4.2%
Expert Review	29.2%	Play With Layer	4.2%
Simulated Users	27.1%	Cognitive Walkthroughs	2.1%
Focus Group	25.0%	Ethnographic Observation	2.1
Cross-Validation	22.9%	Cooperative Evaluation	2.1%

Other evaluation methods included: 6.3% (eye-tracking, task completion time, system preference survey), Latin squares, formative and summative evaluation (Ainsworth et al.,

1999; Barros, 1999; Guzmán, 2005; Mark and Greer, 1993; Shute and Regian, 1993). The Most commonly reported methods used were questionnaires followed by experimental observations, interviews and user observation, respectively.

The population standard deviation for the dataset of evaluation methods was computed using the following formula:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - N)^2} = 9.56$$

Where:

$\sigma$  = population standard deviation

$x_1 \dots, x_N$  = evaluation methods dataset

$\mu$  = mean of the evaluation methods population dataset

$N$  = size of the evaluation methods population dataset

Dataset(evaluation methods depicted in Table 5-2) = (24, 36, 12, 3, 24, 2, 4, 20, 2, 13, 11, 7, 9, 14, 0, 1, 0, 7, 4, 22, 29, 0, 17, 7, 17, 1, 15, 3, 1, 3)

Total number = 30

The results after computing the mean (average) and standard deviation:

Total numbers:	30
Mean (average):	10.27
Standard deviation:	9.72

### ***Measurement Criteria (Adaptive Variables)***

It is important to ensure that the correct measurement criteria and metrics are used with the correct evaluation method. A total of 43 measurement criteria – also known as adaptive variables – were reported. Table 5-3 presents a summary of these results.

**Table 5- 3: Reported evaluation criteria (factors)**

<b>Criteria (factors)</b>	<b>Percentage 75% to 21%</b>	<b>Criteria (factors)</b>	<b>Percentage 20% to 0%</b>
Usability	69.6%	Early Prototype Evaluations	21.7%
Perceived Usefulness	65.2%	User Goal	(19.6%)
User Satisfaction	65.2%	Trust and Privacy Issues	(17.4%)
User Performance	54.3%	Interface Knowledge	(13.0%)
User Behaviour	47.8%	To Combine Qualitative Evaluation	(15.2%)
Intention to Use	39.1%	Collaboration with Real Users During Final Evaluation Step	13.0%
Appropriateness of Adaptation	38.4%	Transparency	10.9%
Usability of Interface Adaptation	30.4%	Appropriateness	8.7%
Content Adaptation	28.3%	User Cognitive Workload	8.7%
Preferences	26.1%	Background and Hyperspace Experience	6.5%
Knowledge of Domain	23.9%	Hyperspace Experience	4.3%

Others (13.0%) included: Contents reutilization capabilities, syntactic and semantic interoperability, effectiveness (for decision support), piloting, precision and recall, and system response time. The most commonly used were usability, user satisfaction, perceived usefulness, and user performance respectively.

In addition, also computed were; mean (average), standard deviation, variance, population standard deviation and population standard deviation for the dataset of evaluation criteria(s).

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} = 8.95$$

Where;

$\sigma$  = population standard deviation

$x_1 \dots, x_N$  = evaluation criteria dataset

$\mu$  = mean of the evaluation criteria population dataset

$N$  = size of the evaluation criteria population dataset

Dataset (evaluation criteria depicted in Table 5-3) = (32, 30, 18, 8, 16, 22, 9, 11, 3, 14, 30, 6, 2, 10, 4, 13, 12, 11, 25, 5, 4, 4, 13, 7, 0, 0, 6, 6)

Total number = 28

The results after computing the mean (average) and standard deviation:

Total numbers:	28
Mean (average):	11.46
Standard deviation:	9.11

### ***Evaluation Metrics Used***

A total of 32 metrics were reported. Mostly commonly used metrics were accuracy of recommendations, precision, accuracy of retrieval, and reliability metrics. A summary of results is shown in Table 5-4.

**Table 5- 4: Reported measurement metrics**

<b>Metrics</b>	<b>Percentage 63% to 2%</b>
Accuracy of Recommendations	62.5%
Precision	59.4%
Accuracy of Retrieval	37.5%
Reliability Metrics	18.8%
Behavioural Complexity	12.5%
pIA: Performance Influence on Adaptivity	12.5%
pQoR: Performance Quality on Response	9.4%
Personalization Overall Cost	9.4%
ApOC: Adaptive Personalization Overall Cost	9.4%
pLatency: Performance Latency	6.3%

**Table 5- 4: Continued**

<b>Metrics</b>	<b>Percentage 63% to 2%</b>
UiAI: User Interaction Adaptivity Index	6.3%
AvgpACF: Average Personalization Adaptive Cost Per Functionality	3.1%
Software Size and Length Metrics	3.1%
MpAC: Minimum Personalization Adaptive Cost	3.1%
DSAI: Domain Specific Adaptivity Index	3.1%

Other metrics reported include: ApOC: Adaptive, Others (37.5%), which included task completion time, task effectiveness, task efficiency, whether users continued to use the system, perceived appropriateness of adaptations, invasiveness of adaptations, awareness of adaptations, response time, time between user request and presentation of the response, SUS-specific questionnaire that was built for the project, Mean Reciprocal Rank (MRR) and Success@K (= probability that relevant item occurs within the top k of the recommendation ranking), user satisfaction, and use of recommended items.

The population standard deviation for the dataset of evaluation metric(s) was computed by:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - N)^2} = 6.18$$

Where:

$\sigma$  = population standard deviation

$x_1 \dots, x_N$  = evaluation metric(s) dataset

$\mu$  = mean of the evaluation metric(s) population dataset

N= size of the evaluation metric(s) population dataset

Dataset (measurement metrics depicted in Table 5-4) = (20, 12, 0, 4, 6, 19, 1, 2, 2, 3, 4, 1, 1, 0, 0, 3, 1, 12)

Total number = 18



The results after computing the mean (average) and standard deviation:

Total numbers:	18
Mean (average):	5.06
Standard deviation:	6.36

It is important that evaluators understand the internal models of adaptive systems. The next section presents a summary of these models.

### 5.2.3.3 Internal Models of Adaptive Systems

To find out how adaptive systems had been developed, participants were presented with a list of internal models which the candidate had identified when conducting the literature review, and asked them to ‘Please tick the meta data models your system used’.

When asked to name meta data models used when developing the adaptive systems , majority of the participants stated they had used the user model (90.6%), followed by the content model (50.9%), domain model (45.3%), presentation model (24.5%), navigation model (24.5%), device model (9.4%), , task model (18.9%), strategy model (17.0%), system model (3.8%), and other (20.8%) which included: group model, evidence model, affective model, adaptation history model, game state model, virtual patient model, bug model, scenario model, educational standards (IMS family), and adaptation model.

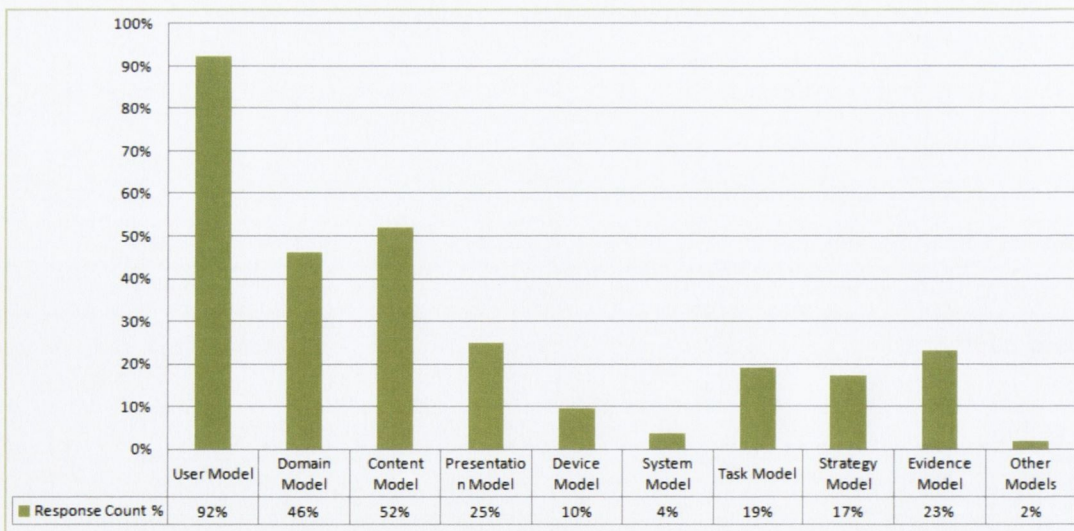


Figure 5- 2: Internal models used when developing an adaptive system

Computed mean and standard deviation is:

Dataset (internal models, depicted in Figure 5-2) = (48, 24, 27, 13, 5, 2, 10, 9, 13, 10)

Total number = 10

The population standard deviation for the dataset of evaluation metric(s) was computed by:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - N)^2} = 12.90$$

The results after computing the mean (average) and standard deviation:

Total numbers:	10.0
Mean (average):	16.10
Standard deviation:	13.60

The most commonly used metadata model of adaptive systems was the user model, followed by the content and domain models. A summary of the reported models is shown in Figure 5-2. The results provide evidence that there were limited evaluations of such models.

#### 5.2.3.4 Evaluations of the Internal Models of Adaptive Systems

The reporting of evaluations of the internal models was poor. Only nine participants responded when asked: ‘If you conducted evaluations of specific internal models of adaptive system, what evaluation methods did you use (i.e. for each model evaluated, please indicate which evaluation methods and criteria you used)?’ Of these, one stated that they did not evaluate the models and two stated that they did not understand the question.

Table 5-5 presents the response of each participant.

**Table 5- 5: Response from participants on whether they had evaluated the internal models**

1	Personally, I mostly worked on the student model, which I did evaluate using simulated students (for initial calibration etc.) – later I did evaluate the model using real interaction data, which could be "replayed" into the system. The main focus here was in determining the quality of predictions about exercise results based on the current information about a student.
---	--

		9/19/2011 8:44 PM
2	In one example, we evaluated a Bayesian student model that assessed students' knowledge during gameplay. I performed a statistical comparison of the student model's estimates to a computer-based post-test. Part of this work was published in the following paper: Jonathan Rowe and James Lester. Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments. In Proceedings of the sixth annual AI and interactive digital entertainment conference, Palo Alto, California, pp. 57-62, 2010.	9/1/2011 7:07 PM
3	We used randomized controlled evaluations wherever possible; see this for several examples: <a href="http://teacherwiki.assistment.org/wiki/Publications">http://teacherwiki.assistment.org/wiki/Publications</a>	9/1/2011 1:53 PM
4	calculating precision and recall	9/1/2011 9:53 AM
5	no	9/1/2011 8:31
6	I do not understand the question	9/1/2011 7:12 AM
7	I don't understand "internal models". Generally we used different techniques according to our goal and as much of my work aimed to provide scrutable interfaces, usability was always a key issue and we used many approaches to assessing that.	8/31/2011 11:44 PM
8	We evaluated the user model and tested different user modeling strategies, i.e. we just switched user modeling and checked with what strategy we gain the best recommendations (= overall output of the system). We did two studies: (1) live online evaluation (duration: 1 month): here we switched between 4 UM strategies; at the end of the evaluation we analyzed with which UM strategy we produced most "I like the recommendation" ratings. (2) offline evaluation (cross-validation): here we tested further strategies and tried to optimize the recommendations by tweaking the UM strategies	8/31/2011 6:09 PM
9	Comparison between adaptive and linear versions of systems in terms of learning gains and user satisfaction.	9/13/2011 9:44 PM

Although the focus of this research is not on how researchers had evaluated internal models, the findings show that there is a gap in evaluations of internal models' adaptive systems. In our opinion, this is an area that future researchers should focus on.

The educational evaluation dataset resulting from this study was added to the one already collected during the literature view (chapter 2). The two datasets added together were used to populate the database of the EFEx evaluation framework (discussed in Chapter 4). The next section presents a summary of this dataset.

### 5.3 Author Contribution: Educational Evaluation Dataset

Datasets for educational adaptive E-Learning (AEL) are manifold as AEL takes place in the whole spectrum of learning, roughly distinguished between formal and non-formal learning settings. Although AEL systems are increasingly applied in E-Learning, it is still an application area that lacks publicly available, comparable, interoperable and reusable datasets that cover the spectrum of formal and informal learning.

The individual contribution of the author is an evaluation educational dataset for adaptive systems which the candidate has been collecting over a period of five years (Mulwa et al., 2011). A review and analyses of a total of 350 evaluation studies of adaptive systems were conducted, over a period of five years, more specifically focusing on evaluations of AEL systems. Based on the analysis, an evaluation educational dataset for supporting evaluators of such systems was created. The dataset is broken down into eight distinct interlinked structures; depicted in Table 5-6.

**Table 5- 6: Summary of the educational dataset**

<b>Dataset</b>	<b>Total</b>
Evaluation of adaptive systems and their internal models studies	80
Adaptive systems	105
Categories of adaptive systems	13
Evaluation criteria (factors)	75
Evaluation methods	74
Measurement metrics	85
Internal models of adaptive systems	15
Evaluation approaches	7

Furthermore the results were characterized, structured and interlinked to form a list of evaluation techniques. Table 5-7 presents a subset of these dataset.

**Table 5- 7: Subset of the educational evaluation dataset**

<b>Evaluation Method/ Instrument</b>	<b>Criteria (Variables)</b>	<b>Metrics</b>	<b>References</b>
Interviews, questionnaires (online, post-test, pre-post-test, verbal), focus group, discussion groups.	Usability, perceived usefulness, Intention to use, user goals, knowledge of the domain, background and hyperspace experience, preferences trust and privacy issues, appropriateness of adaptation.	Accuracy of recommendations, accuracy of retrieval, AiAI: administrator interaction adaptivity index.	(Gena, 2005c), (Van Velsen et al., 2008), (Masthoff, 2006, Raibulet and Masciadri, 2009) .
User observation, systematic observation, verbal protocol, data mining, play with layer, simulated users, Cross-validation, heuristic evaluation.	Usability, user behaviour, user goal, knowledge of domain, background and hyperspace experience, and user interests individual traits (e.g. cognitive or learning style), environment (e.g. location, locale, software, and hardware), and user situation awareness.	behavioural complexity, reliability metrics, precision, software size and length metrics UiAI: user interaction adaptivity index.	(Gupta and Grover, 2004), Rothock et al. 2002, (Magoulas & Dimakopoulos, 2005), Steehouder M. 2008, (Brusilovsky, 2001) .
Heuristic evaluation, expert review, parallel design, cognitive walkthroughs, social-technical models.	Usability of interface adaptation & user, domain and interface knowledge, user performance.	pQoR: performance quality of response.	Rothock et al. 2002.

Table 5-7 Continued

Evaluation Method/ Instrument	Criteria (Variables)	Metrics	References
Wizard of Oz simulation, scenario-based design, prototypes.	Early prototype evaluations, evaluation before implementation.	pIA: performance influence on adaptivity.	Judith Masthoff, 2006.
Usability testing, experimental evaluation.	Interface (and content) adaptation, usage data (user history), user cognitive workload, groups of users.	MpAC: minimum personalization adaptive cost.	Magoulas & Demakopoulos, Rothock et al. 2002.
Cultural probes, focus group, user-as-wizard heuristic evaluation, cognitive walkthrough, simulated users, play with layer, user test.	Preferences, user interests, user skills and capabilities, user performance.	AvgpACF: Average personalization adaptive cost per functionality.	(Santos, 2008), (Masthoff, 2006), (Paramythis et al., 2010)
Empirical observations, heuristic evaluation, cognitive walkthrough, user test, play with layer.	User cognitive workload, appreciation, trust and privacy issues, user experience.	pOCF: personalization overall cost per functionality.	(Díaz et al., 2008)
Creative brainstorming sessions, focus group, user-as-wizard.	Privacy, transparency, appropriateness, appreciation, trust and privacy issues, user experience, user satisfaction, usability, user behaviour, intention to use, perceived usefulness.	MpOCF: minimum personalization overall cost.	(Van Velsen et al., 2008)
Questionnaire, interviews, ethnographic observation,	Real user actions, user behaviour, intention to use, perceived usefulness	AvgpACF: average personalization overall cost per functionality	Gena 2005, Diaz et al. 2008, Gena 2005

**Table 5-7: Continued**

<b>Evaluation Method/ Instrument</b>	<b>Criteria (Variables)</b>	<b>Metrics</b>	<b>References</b>
Quantitative, grounded theory, cognitive walkthrough, heuristic evaluation, user test.	To combine qualitative evaluation, to discover new theories.	ApOC: adaptive personalization overall cost.	Diaz et al. 2008, Gena 2005
Prototyping, heuristic evaluation, cognitive walkthrough, user test, play with layer, cooperative evaluation, verbal protocols, and focus group.	Evaluation of vertical or horizontal prototype, Collaboration with real users during final evaluation step.	DSAI: domain specific adaptivity index.	Gena 2005

Following is a brief discussion of the evaluation metrics and approaches:

*Evaluation Method/Instrument:* From the analysed studies; questionnaires, experimental evaluation, interviews, user observations, usability testing were the most commonly used evaluation methods respectively. Questionnaires were used to collect data from respondents by allowing them to answer a set of questions either on paper or online. Participants could choose one or multiple choices or can answer freely in writing.

Current evaluation approaches recommend experimental methods (techniques) in lab settings as a way of coping with the complexity of adaptive systems and identifying the aspects of these systems that require improvement. In interviews (structured, fixed questions, or semi-structured), participants normally are asked in person by an interviewer. The manner in which interview results were reported indicates that evaluators considered interviews to be inferior to statistical data. Usability testing methods are used in user-centered interaction design to evaluate a system by testing it on users. This focuses on measuring the system's 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 the other 25 were identified from the literature. These variables were then grouped into different categories (i.e. attitude and experience, actual use, system adoption and system output). *Usability* was most frequently measured, followed by *User Satisfaction*, a subjective variable which can be influenced by various factors (such as system effectiveness, user effectiveness, user characteristics, effort and effectiveness. Keinonen defines usability as “a characteristic related to: i) the product’s design process, ii) the product itself, iii) use of product, iv) user experience of the product and user expectation” (Nokelainen, 2006). These are attributes which can be measured through subjective user experience.

*Metrics:* In the analysed studies, accuracy of the recommendations metric was the most frequently used, followed by accuracy of retrieval. Furthermore, evaluation approaches reported included quality approach, lifecycle approach, combined and layered evaluation approach, combined four-level and six-level approach, user-centered evaluation approach, empirical approach, utility approach, collaborative filtering, content-based, demographic, knowledge-based, and hybrid (Ehlers et al., 2005, Drachsler et al., 2010, Breitner and Hoppe, 2005, Mulwa c. et al., 2011).

## 5.4 Conclusion

The results of this experiment enabled us to partially tackle **Sub RQ1** (formulated in Section 1.2): ‘*What are the techniques used and tradeoffs between those techniques to support user-centered evaluations of adaptive systems?*’

It is crucial that software developers and evaluators evade well-known pitfalls and that writer of future evaluation reports increase their empirical value, by reporting the approaches used. In this study, evaluation approaches are considered as any technique, 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.

The candidate believes that the dataset in this study is the first harvested dataset of selections of adaptive evaluation approaches, methods, metrics and criteria for AEL. The key aspect is that this information does not consist of arbitrary selections from novice end users but of peer-reviewed informed choices from published researchers. Although recommendation systems can be applied to large datasets, this does not mean that



recommendation systems are inappropriate or are not needed to solve complex information problems, and, in particular, the multi-attribute relationships which need to be traversed to work out what are the most appropriate evaluation procedures. Evaluation approaches, methods/techniques, metrics and criteria are not easily navigated using typical database techniques.

Furthermore, the educational dataset created as a result of this research will provide the much-required educational evaluation data for evaluation of adaptive systems. Currently the reporting of UCE studies is poorly conducted and no data exists on how different models for adaptive systems have been evaluated. Provision of the collected data in a structured way will encourage research in this application area.

# Chapter 6: Evaluating EFEx – Recommendations Accuracy, Search Identification and Taxonomy Usability

## 6.1 Introduction

The objective of this chapter is to evaluate the three components of EFEx framework as previously defined in chapter 1.

**Objective4**      *Evaluate the three components of the evaluation framework designed in Objective3:*

- (i) An automated hybrid (case-based and knowledge-based) system for recommending evaluation techniques,*
- (ii) A personalised search component that allows users to find evaluation studies of adaptive systems and*
- (iii) A taxonomy of technical terms for supporting the evaluation of adaptive systems.*

In order to evaluate EFEx framework; three evaluation experiments were conducted. The first experiment targeted the hybrid recommendation system. In particular the novice evaluators<sup>26</sup> were asked to:

**Experiment 1 (Novice evaluators):**

- Objective 1: Did the recommender identify appropriate evaluation techniques.
- Objective 2: Were the novice evaluators satisfied (and able to learn) after interaction with the recommender system.

**Experiment 2 (Expert evaluators):**

The expert evaluators<sup>27</sup> were asked to evaluate the method choice for adaptive systems evaluations, as such, is the appropriateness of explanation of the technique.

---

<sup>26</sup> Novice evaluators are developers of adaptive systems (2000-2013)

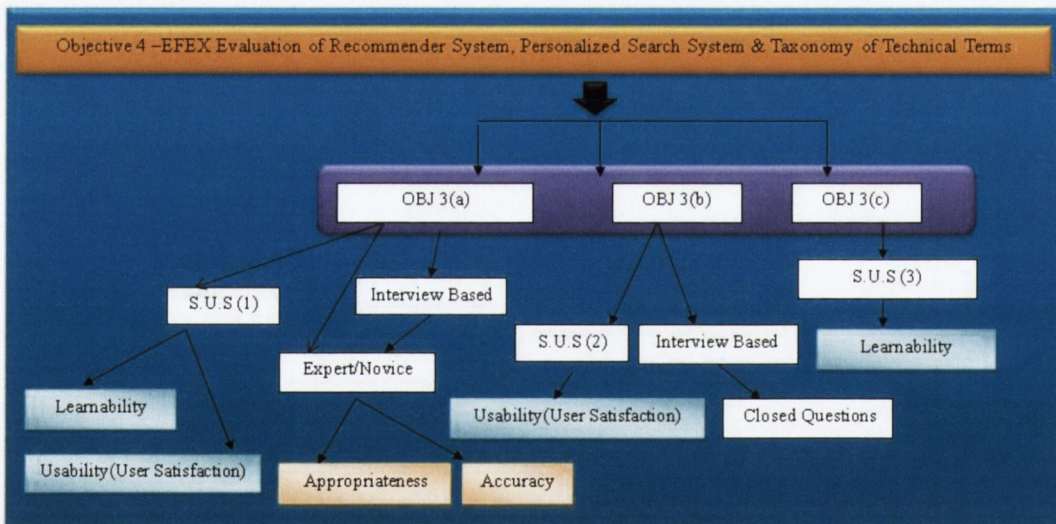
<sup>27</sup> Expert evaluators are developers of adaptive systems ((2000-2013)

The second experiment targeted the personalised search sub system. In particular the novice evaluators<sup>28</sup> were asked to:

- Objective 1: Were the novice evaluators able to identify evaluation studies: (i) evaluation studies of internal models of adaptive systems, (ii) evaluation studies of adaptive systems and (iii) general evaluation studies of adaptive systems.
- Objective 2: Were the novice evaluators satisfied (and able to learn) after interaction with the personalised search system.

The third experiment aimed at identifying users' appreciation and satisfaction regarding the various functionalities provided by the taxonomy of technical terms of evaluation of adaptive systems.

Several evaluation methods were used in the three experiments: (i) case-study using a real-life user-study on what people are doing in evaluation of adaptive systems, (ii) use of automated recommendation techniques, to test the appropriateness of recommendations, (iii) structured interview-based technique, and (iv) system usability scale questionnaires (i.e. user satisfaction and learnability). The process of evaluating research objective 2 and 3 is depicted in Figure 6-1.



**Figure 6-1: Evaluation objectives of EFEX framework**

Overall the evaluation seeks to evaluate the appropriateness of recommendations (evaluation techniques and approaches) that the hybrid recommender system produces. In

<sup>28</sup> Novice evaluators are developers of adaptive systems (2000-2013)

addition it evaluates the usability (perceived usefulness and learnability) of the; (i) hybrid recommender system, (ii) personalised search sub system and (ii) taxonomy of technical terms.

### ***6.1.1 Chapter Organization and Objective***

This chapter is structured as follows:

Section 6.2 presents the evaluation objectives of the recommender system and presents the results and findings of expert and novice evaluators. Section 6.3 presents the evaluation objectives of the personalised search sub-system. It also presents the results and findings of novice evaluators after interacting with the system are presented.

Furthermore Section 6.4 presents the evaluation objectives of the taxonomy of technical terms. It also presents the results and findings after novice evaluators have interacted with the taxonomy. Finally Section 6.5 concludes the chapter.

## **6.2 A Hybrid Recommender System: Recommendation Appropriateness**

The presented hybrid recommender system was evaluated in terms of (i) accuracy and appropriateness of recommendations (Manouselis et al., 2011) (ii) usability (user satisfaction and learnability) and (iii) educational benefits. This section presents the results and findings of a task-based user trial of the recommender system presented in section 4.4.1 above.

### **6.2.1 Experiment for Novices**

The experiment is divided into two parts. First the novice evaluators are given the recommender system to use. The main aim is to see if: (i) they are convinced by it, (ii) whether terms of usability; they perceive it to be useful and are able to learn. In the second part, we are directing the experts through how they would choose evaluation techniques and seeing for generic systems what techniques the experts would recommend. The aim is to find out under what conditions the experts would recommend evaluation techniques and

approaches. The next step in the evaluation involves comparing the results of the experts to what the hybrid recommender system would produce for specific systems.

### **6.2.1.1 Experiment Objectives**

The goal of this experiment was to find out if the novice evaluators after interacting with the hybrid recommender system were able to effectively identify appropriate evaluation techniques. In addition the experiment aimed at finding out perceived usability and learnability (i.e. were the evaluators able to learn after interacting with the recommender system). The two objectives formulated in this experiment are:

#### ***Objective 1: Identification of Appropriate Techniques:***

The first objective aimed at finding out whether recommender identified appropriate evaluation techniques:

- Evaluation methods to be used when evaluating an adaptive system?
- Measurement criteria to be used when evaluating an adaptive system?
- Evaluation metrics to be used when evaluating an adaptive system?
- Evaluation approaches to be used when evaluating adaptive systems

In addition, the experiment aimed at finding out what features (i.e. characteristics) of the recommended evaluation techniques did the novice evaluators like most (find useful) about the recommender system?

#### ***Objective 2: Usability (User satisfaction and Learnability)***

In terms of user satisfaction, the benefit to users lies in the perceived usability of the various recommendation functions provided by the hybrid recommender system. User satisfaction is typically measured through usability questionnaires after completing given tasks with a system. This experiment objective aimed at finding out:

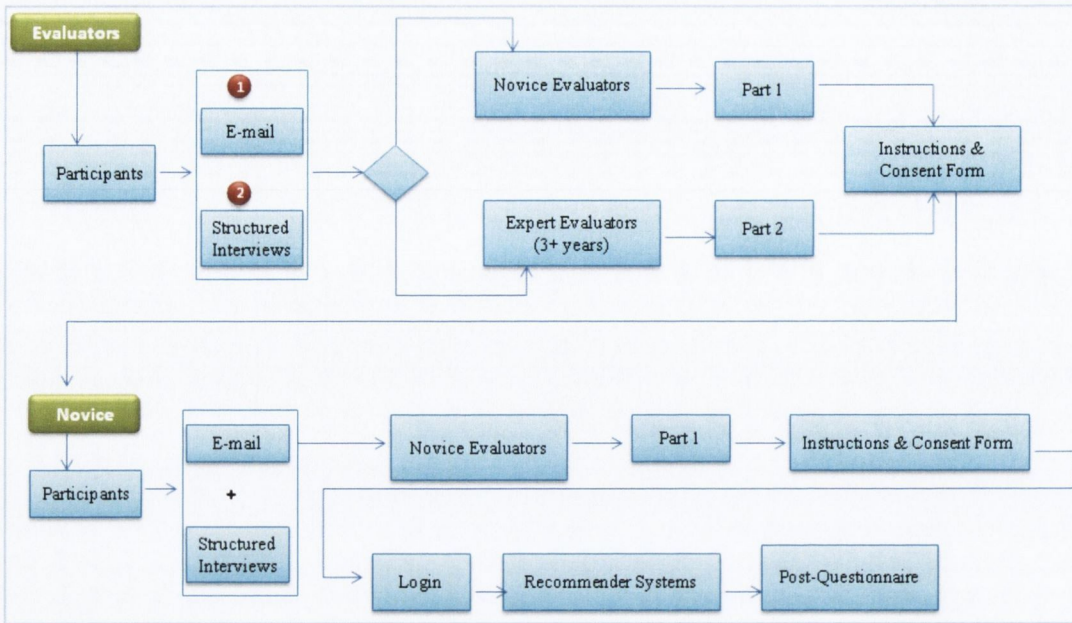
- Were the novice evaluators satisfied after interaction with the recommender system
- Were novice evaluators satisfied (and able to learn) after interaction with the recommender system.

In this thesis learnability describes the ability of the recommender interface to allow users to accomplish tasks on the first attempt. In most cases it is often referred to as usability for first-time use. Nielsen also defines learnability as easy first-time use but lists it as a sub-component of the construct of usability. In addition the experiment aimed at finding out if evaluators would *not need the support* of a technical person to be able to use the recommender system or *not need to learn* a lot of things before they could get going with the recommender system. In order to test this objective usability scores (of Q4 and Q10 of SUS questionnaire) are used.

#### **6.2.1.2 Experiment Setup**

A total of 53 novice evaluators participated in the online and structured interview-based experiment. Of these, 43 completed the full evaluation process. Participants were recruited from Trinity College Dublin, Dublin Institute of Technology, the UMAP (2011, 2012 and 2013) conference and the AH conference, DataTEL and Recommender Communities.

Each evaluator received an email about the purpose and duration of the experiment, as well a URL link to the experiment and the recommender system. The participants were then asked to fill out a consent form (see Appendix D3.1) in order to determine their expertise in evaluation of adaptive systems. The structured-interviews participants were given the exact same questions which had been given to the online participants. The whole experimental process is depicted in Figure 6-2.



**Figure 6-2: Experimental Process for Novice Evaluators**

Before tackling evaluation objectives one and two, i.e.:

- *Objective 1: Did the recommender identify appropriate evaluation techniques.*
- *Objective 2: Were the novice evaluators satisfied (and able to learn) after interaction with the recommender system.*

Participants were informed they will be interacting with an automated hybrid (case-based and knowledge-based) recommender system. This system will recommend to you the most appropriate evaluation approach and techniques (methods, metrics and criteria) for evaluating an adaptive system. The system will also recommend bundles (combination of) the most appropriate (method + measurement criteria and metrics) which can be used together during evaluation of such a system. Throughout the recommendation process, you will be provided with explanations as to how the recommended techniques were derived.” Then next you will be asked to complete a system usability scale (SUS) questionnaire which has a Likert scale of 1 (completely disagree), 2 (somehow disagree), 3 (somehow agree), 4 (agree) and 5 (strongly agree), including the standard Usability Scale (SUS).

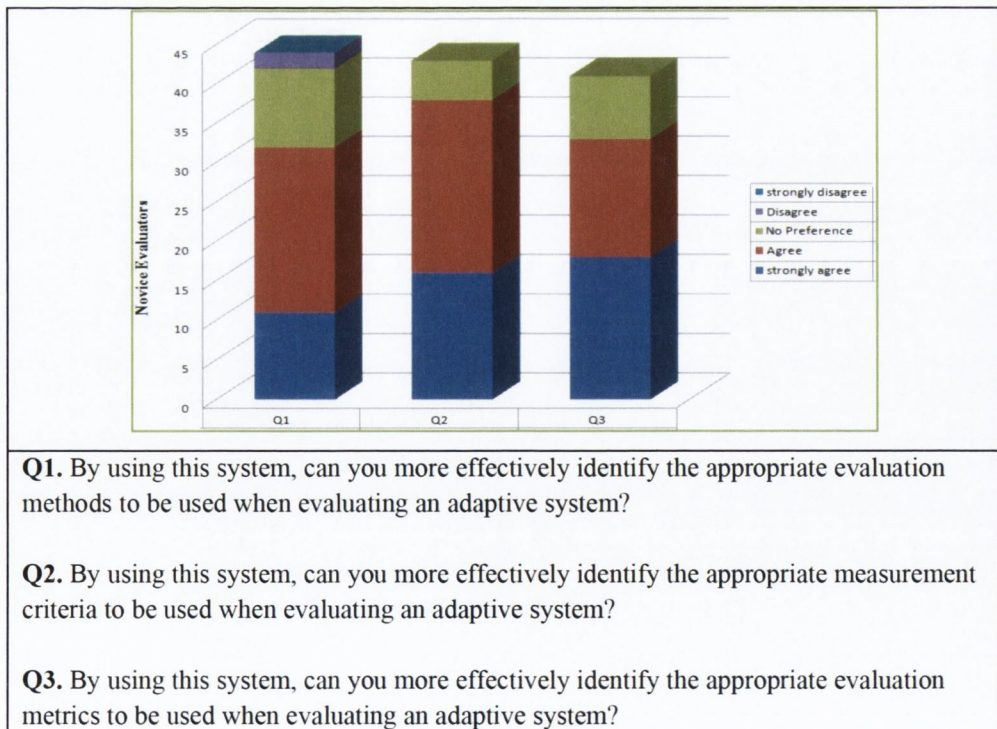
### 6.2.1.3 Evaluation Results and Findings

This section presents the results and findings of experiment 1.

#### *Appropriateness*

When asked to respond to the question “By using this system, can you more effectively identify the appropriate evaluation methods to be used when evaluating an adaptive system?” 79.07% agreed while 4.60% disagreed. When asked to respond to the question “By using this system, can you more effectively identify the appropriate measurement criteria to be used when evaluating an adaptive system?” 88.37% agreed while 2.32% disagreed.

Furthermore, when asked to respond to the question “By using this system, can you more effectively identify the appropriate evaluation metrics to be used when evaluating an adaptive system?” 76.4% agreed while 4.6% disagreed. Figure 6-11 presents a summary of these results. These results are very encouraging for such a novel system, especially considering that most users had not previously used an educational hybrid recommender system.



**Figure 6- 3: Identification of appropriate evaluation techniques**



## *System Usability Scale (SUS)*

Since its introduction in 1986, the 10-item System Usability Scale (SUS) has been assumed to be unidimensional (Nielsen, 1994). Factor analysis of two independent SUS data sets reveals that the SUS actually has two factors, namely Usability (8 items) and Learnability (2 items). These new scales have reasonable reliability (coefficient alpha of .91 and .70, respectively). They correlate highly with the overall SUS ( $r = .985$  and  $.784$ , respectively) and correlate significantly with one another ( $r = .664$ ), but at a low enough level to use as separate scales (Lewis and Sauro, 2009). While SUS was only intended to measure perceived ease-of-use (a single dimension), recent research by Lewis (2009) shows that SUS provides a global measure of system satisfaction and subscales of usability and learnability. Questions 4 and 10 provide the learnability dimension and the other eight questions provide the usability dimension. This means you can track and report on both subscales and the global SUS score.

In addition to the task-based questions, the user study aimed at identifying users' appreciation and satisfaction regarding the various functionalities provided by the recommender system. Out of the 43 users who participated in this study, 31 completed the SUS questionnaire.

First of all, in order to determine the overall usability, standard usability scale (SUS) scores were calculated for the recommender system. A SUS score above 68 would be considered above average and anything below 68 as below average. To interpret the scores, the candidate converted them to a percentile rank through a process called normalizing.<sup>29</sup> Figure 6-4 shows how the percentile ranks associate with SUS scores and letter grades. It is necessary to score above 80.3 to get an A (the top 10% of scores). This is also the point where users are more likely to be recommending a product to a friend<sup>30</sup>. Scoring at the mean score of 68 results in a C and anything below 51 in an F (the bottom 15%);

---

<sup>29</sup> <http://www.measuringusability.com/sus.php>

<sup>30</sup> <http://www.measuringusability.com/sus.php>

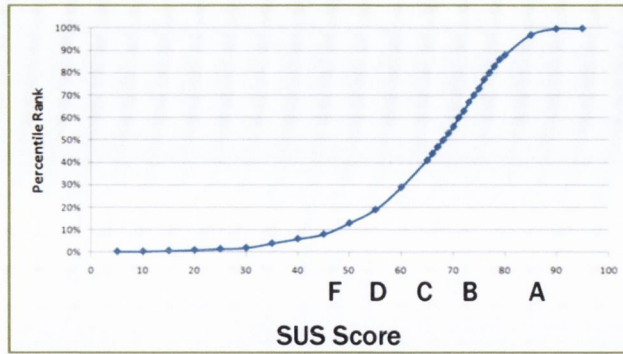


Figure 6-4: Percentile rank associated with SUS scores

In this usability score, the recommender system scored an average of 82%, which is interpreted as a B. This, again, is a very encouraging score for such a novel system, especially considering that most novice evaluators had not used an educational hybrid recommender system in the past. A summary of results of the 31 novice evaluators is depicted in Figure 6-5.

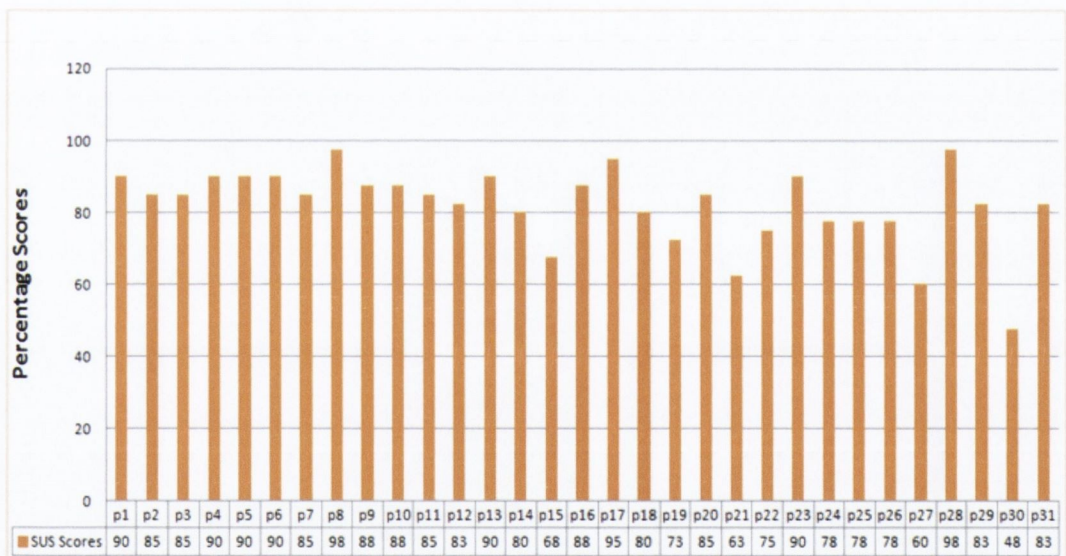


Figure 6-5: Summary of SUS scores by novice evaluators

### User satisfaction

When asked to agree or disagree with the statement “I think that I would like to use this recommender system frequently”, 61% strongly agreed that they would use the recommender system frequently while a small percentage of users 3% strongly disagreed. In addition, when asked to agree or disagree with the statement “I thought the

*recommender system was easy to use*", 68% agreed that the recommender system was easy to use and 3% disagreed.

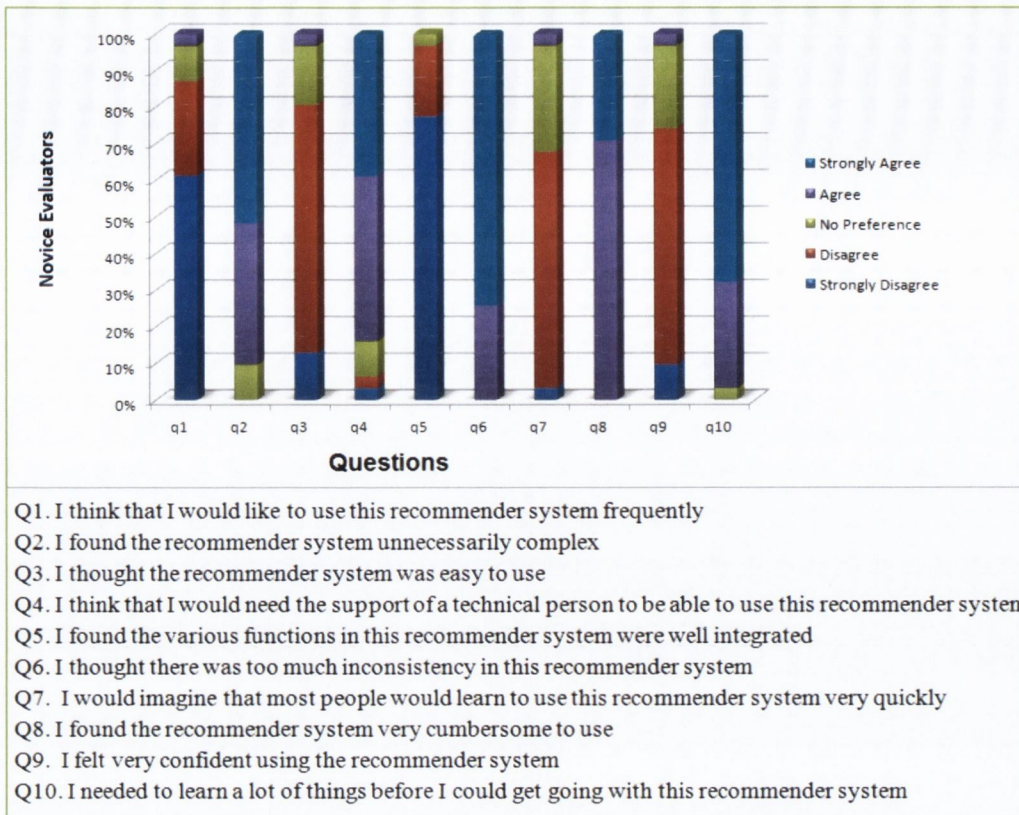
Furthermore, when asked to agree or disagree with the statement "*I found the various functions in this recommender system were well integrated*", 77% strongly agreed that that various functions of the recommender system were well integrated while 3% had no preference. Also, when asked to agree or disagree with the statement "*I would imagine that most people would learn to use this recommender system very quickly*", 65% agreed while 3% disagreed. In addition, when asked to agree or disagree with the statement "*I felt very confident using the recommender system*", 22 out of 31 evaluators (65% agreed) while 3% disagreed

Participants also strongly disagreed or disagreed that the recommender system was unnecessarily complex, had too many inconsistencies and was cumbersome to use. The results are presented below.

When asked to agree or disagree with the statement "*I found the recommender system unnecessarily complex*", 52% strongly disagreed and 39% disagreed while none of the users agreed that they found the recommender system unnecessarily complex.

When asked to agree or disagree with the statement "*I thought there was too much inconsistency in this recommender system*", the majority of evaluators strongly disagreed/disagreed (74%/26%).

Finally, when asked to agree or disagree with the statement "*I found the recommender system very cumbersome to use*", the majority of evaluators strongly disagreed/disagreed (71%/29%) that when interacting with the recommender system they found it cumbersome to use. A summary of these results is presented in Figure 6-6.



**Figure 6-6: Recommender system (user satisfaction and learnability)**

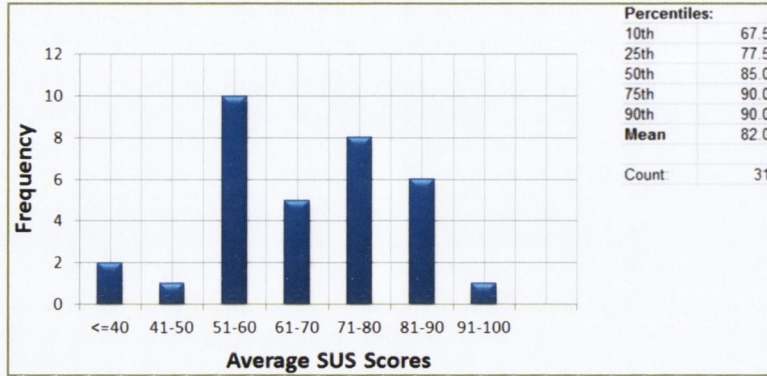
### ***Usability (Learnability)***

Questions 4 and 10 of SUS provide the learnability dimension. The results of users responding to these two questions are depicted in Figure 6-6. When asked to agree or disagree with the statement “*I think that I would need the support of a technical person to be able to use this recommender system*”, the majority of evaluators strongly disagreed/disagreed (45%/39%) while 3% agreed. When asked to agree or disagree with the statement “*I needed to learn a lot of things before I could get going with this recommender system*”, the majority of the evaluators disagreed (68%) while 3% agreed.

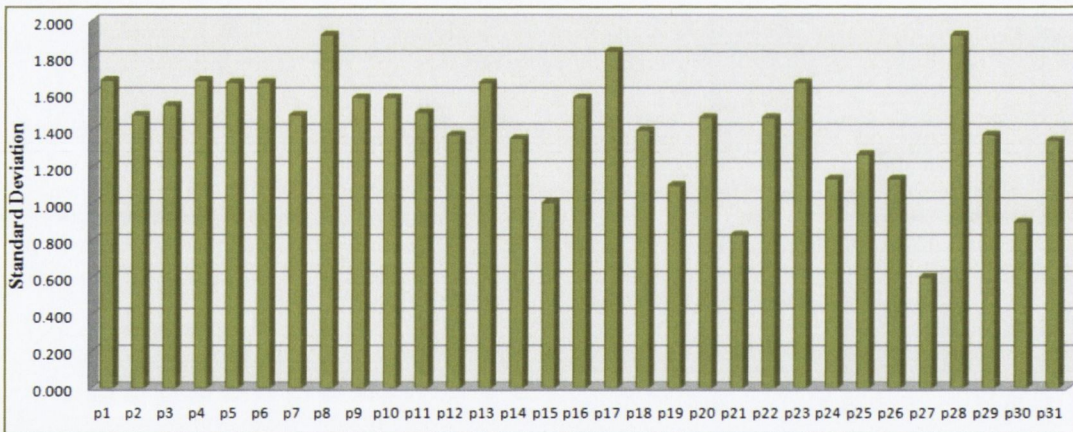
Overall, these are very encouraging results in user satisfaction and learnability for such a novel system, especially considering that most of the users had not used an educational recommender system before.

## Frequency distribution

Due to the large number of data points frequency of distribution was also computed, in order to visualize variability (Figure 6-7).



In order to provide an indication of how far the individual responses to a question vary or deviate, standard deviation (SD) was calculated. Figure 6-8 presents a summary of SD for the 31 participants.



**Figure 6-7: Variance in individual responses to questions, using standard deviation**

In conclusion, based on the results and findings of the novice evaluators after completing experiment 1 (research objective 4), the evaluations of the novices seem to prove that both the evaluations made by the hybrid recommender system seem credible, well argued and well backed up and therefore majority them said the recommended evaluation approaches to be used and techniques seemed to be correct. Unfortunately the candidate cannot

guarantee that the novices are the right people to judge whether the recommendations produced are correct.

## **6.2.2 Experiment for Experts**

### **6.2.2.1 *Experiment Objectives***

Experts were asked to evaluate the method choice for an adaptive systems evaluation, as such, is the appropriateness of explanation of the technique. Ideally what the candidate would like to do is have the experts look at individual systems and then recommend evaluation techniques and approaches to be used and then compare the results with those produced by the hybrid recommender system. However when we tried this most experts were not willing to suggest evaluation for systems they did not develop. Therefore after the discussions with the experts; majority indicated they would recommend evaluation techniques based on variation types of adaptive systems rather than individual system.

### **6.2.2.2 *Experiment Setup***

To participate in this study, users were required to be familiar with adaptive systems or the evaluation of such systems.

A total of 60 expert evaluators participated in the online and structured interview-based study. Of these, 49 completed the full evaluation process. Participants were recruited from Trinity College Dublin, the UMAP (2011, 2012 and 2013) conference and the AH conference, DataTEL and Recommender Communities. This section presents the results of these expert evaluators, who were either very experienced evaluators (3+ years' research experience) or experienced evaluators (1-3 years' experience).

In order to tackle experiment 2, expert evaluators were given a link to the online experiment<sup>31</sup> (see Appendix D) accompanied with instructions to choose a variation type<sup>32</sup> (category) from a list of 13 pre-identified variation types of adaptive systems (see Appendix D3.1.1, Q2). They were also provided with a list of pre-identified adaptive systems belonging to each of the variation types.

---

<sup>31</sup> <https://www.surveymonkey.com/s/G26H7H9>

<sup>32</sup> Variation type (e.g. an adaptive educational hypermedia system)

The experts were then asked to choose properties<sup>33</sup> they wanted to focus on during evaluation and which they felt comfortable recommending evaluation technique(s). Next they were asked which evaluation approach (es) and technique(s) they would recommend to evaluate such an adaptive system(s). To determine which evaluation approach (es) and techniques to recommend, the experts were required to use the properties they had selected. They were also requested to rate the recommendations from 5 to 1 (5 being the most appropriate technique and 1 the least appropriate). The whole experimental process is depicted in Figure 6-9.

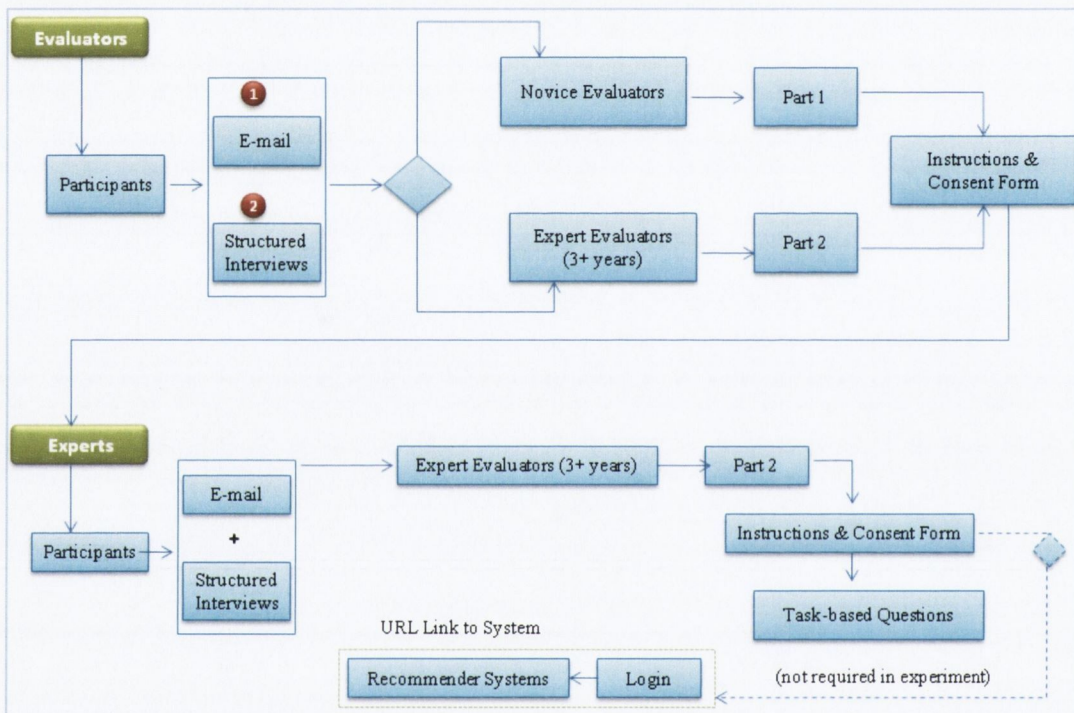


Figure 6- 8: Experimental process for expert evaluators

### 6.2.2.3 Evaluation Results and Findings

This section presents the results and findings of the expert evaluators.

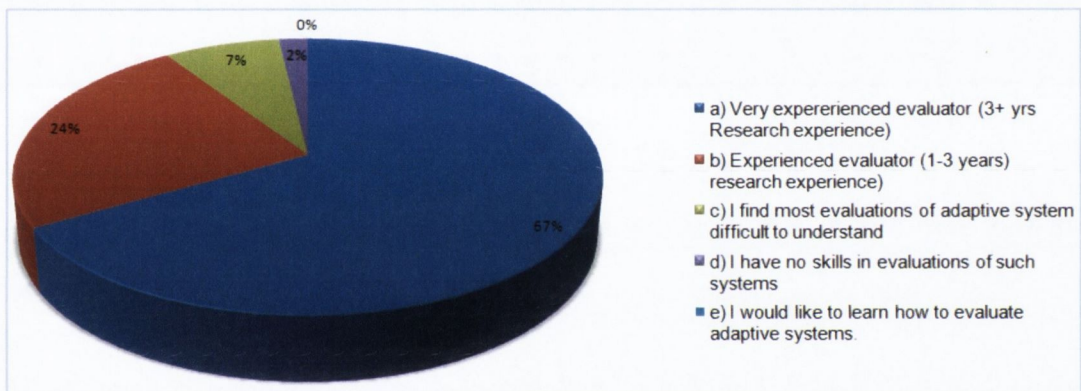
#### User characteristics

In order to capture various user characteristics (i.e. knowledge, experience and expertise in evaluation of adaptive systems), users were asked: “How would you rate your evaluation

<sup>33</sup> An example of a property would be ‘evaluation purpose/goal’

*skills of adaptive systems?”* To answer this question, the participants could select from: (a) Very experienced evaluator (3+ years’ research experience), (b) Experienced evaluator (1-3 years’ research experience), (c) “I find most evaluations of adaptive systems difficult to understand”, (d) “I have no skills in evaluations of such systems”, and (e) “I would like to learn how to evaluate adaptive systems”.

A total of 54 evaluators responded to this question; of these, 67% were very experienced evaluators and 24% had some experience in evaluating adaptive systems. A summary of overall results is presented in Figure 6-10. Although it was specifically stated that only expert evaluators were to participate in this study, 9% of the total participants selected options (c), (d) and (e) above. The results relating to them were disregarded and not included in the analysis.



**Figure 6-9: User characteristics – identification of user experience and expertise of evaluators**

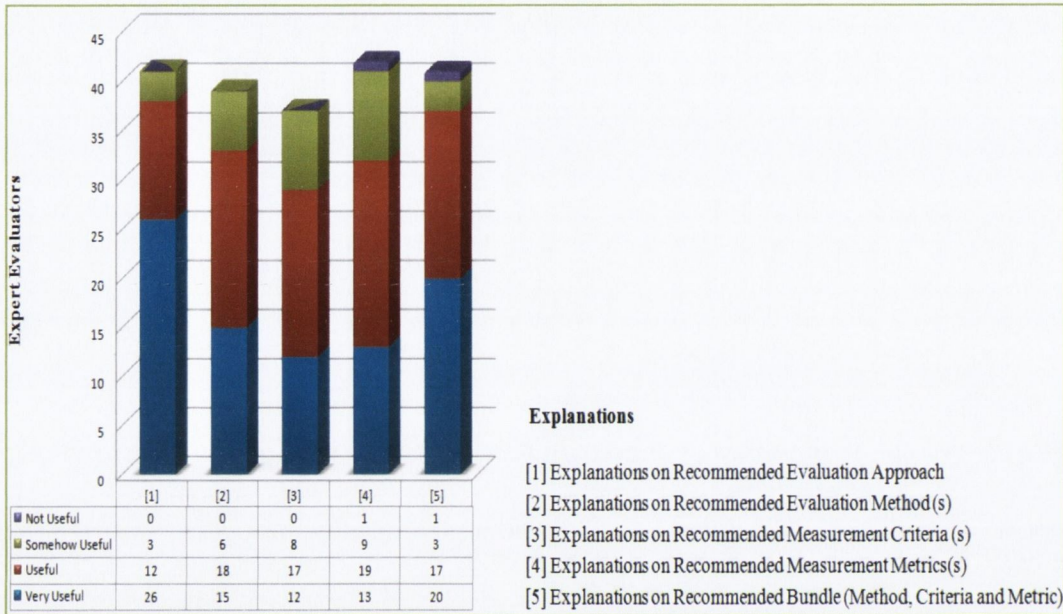
### ***Perceived usefulness***

After identification of user characteristics, first the experiment sought to find out if domain experts in evaluations of adaptive systems perceived such a hybrid recommendation system to be useful in the domain and context of use. When asked “*Do you consider such a system useful?*” 90.9% agreed they would find it useful, while 9.1% disagreed.

In order to gain more insight into which features (provision of explanations regarding recommended techniques) were particularly useful, experts were asked: “*Do you think this explanation would be a useful feature?*” The features rated to be *very useful* were: explanations on recommended evaluation approach (26 users), explanations on recommended bundle (method, criteria and metrics) (20 users), explanations on

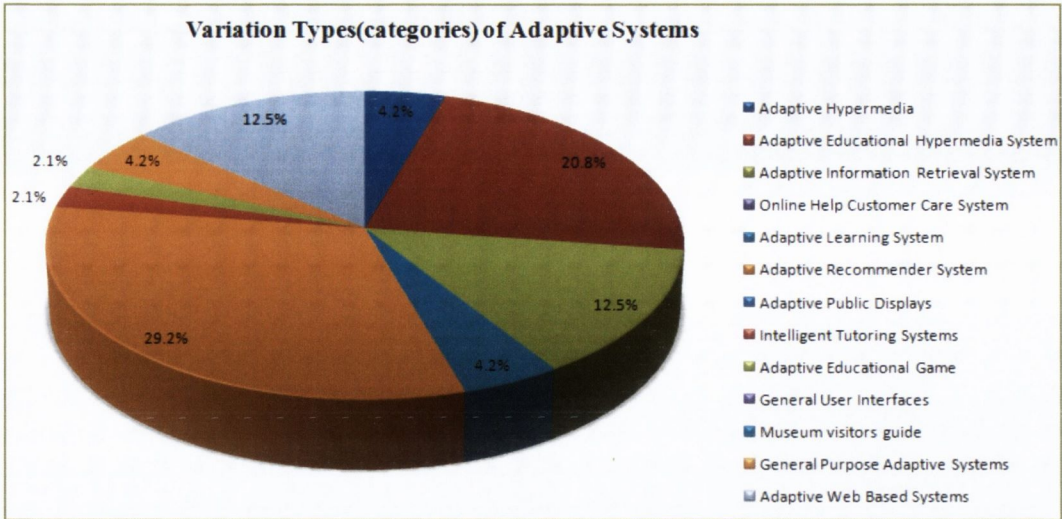


recommended evaluation methods (15 users), explanations on recommended measurement metrics (13 users), and explanations on recommended measurement criteria (12 users). Features rated as being *useful* were: explanations on recommended measurement metrics (19 users) and explanations on recommended evaluation methods (18 users). Only two users rated provision of explanations on the recommended techniques as *not useful*. A summary of all the ratings is presented in Figure 6-11.



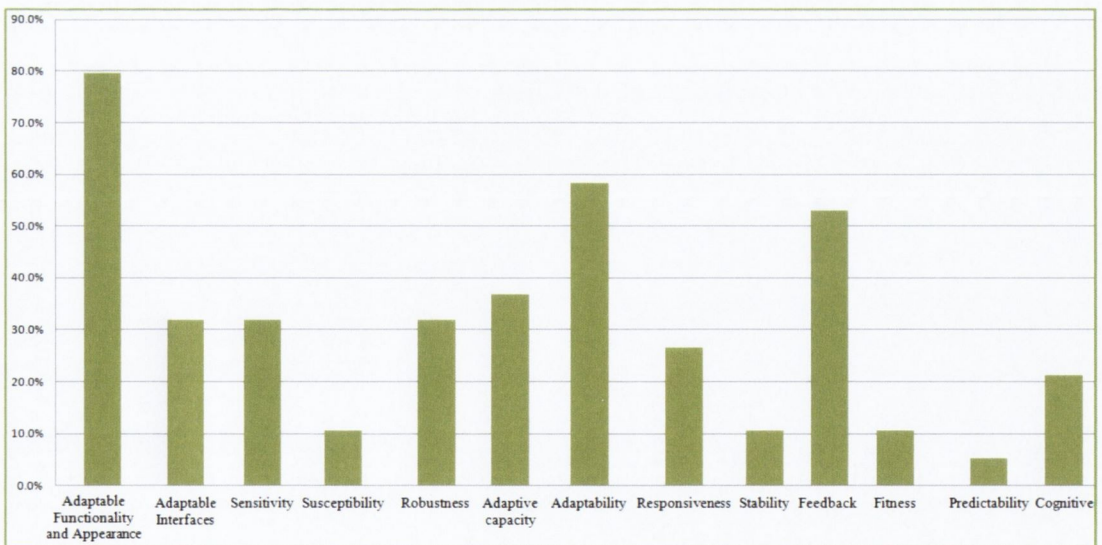
**Figure 6- 10: Explanations on recommended evaluation techniques**

Next, participants were asked to choose “*which category (i.e. also known as variation type) of adaptive systems you wish to focus on during evaluation*”. They were required to choose only one category from a list of pre-identified categories. They were also provided with examples of adaptive systems belonging to the different variation types. A total of 14 out of 48 users selected adaptive recommender systems, while 10 out of 48 selected adaptive educational hypermedia systems. Figure 6-12 presents the overall percentage response.



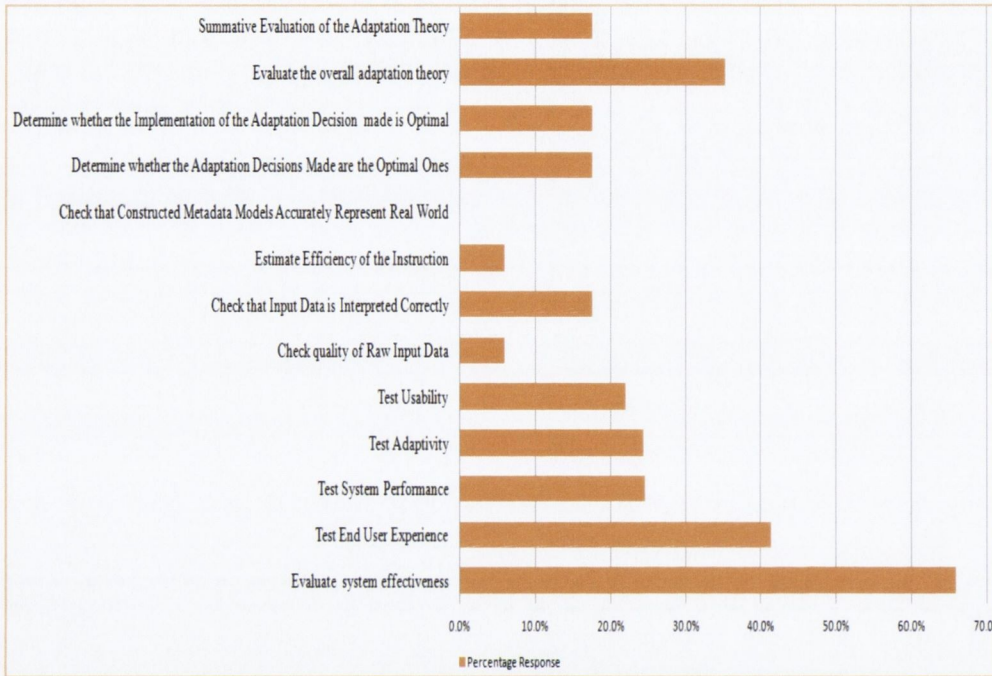
**Figure 6- 11: Types of adaptive systems participants wished to focus on during evaluation**

Before users responded to the next set of questions, they were required to choose which property (or properties) in the case of which they would feel comfortable recommending an evaluation technique (i.e. method, criteria and metrics) to evaluate an adaptive system (i.e. belonging to the variation type chosen in Question Q2). When asked to select “*which system characteristics you wish to focus on during evaluation*”, 79.5% selected adaptable functionality and appearance, 58.3% adaptive capacity and 53.0% user feedback. Figure 6-13 presents a summary of responses on system characteristics.



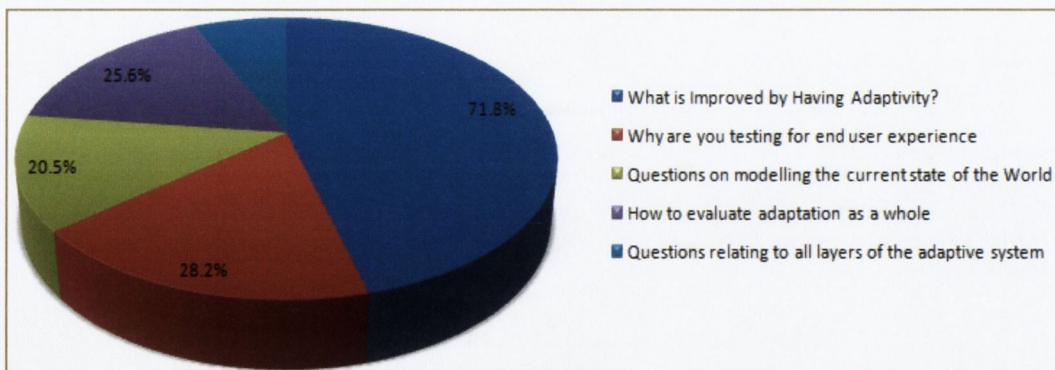
**Figure 6- 12: Experts response on system characteristics**

Based on the categories (variation type) of adaptive system (Q2) and the system characteristics (Q3) that they had selected, the experts were asked: “What would be the goal(s) or purpose(s) of the evaluation being conducted?” In response, 66% wanted to evaluate system effectiveness, followed by 42% who were interested in testing end-user experience. Figure 6-14 depicts a summary of the experts’ evaluation purposes.



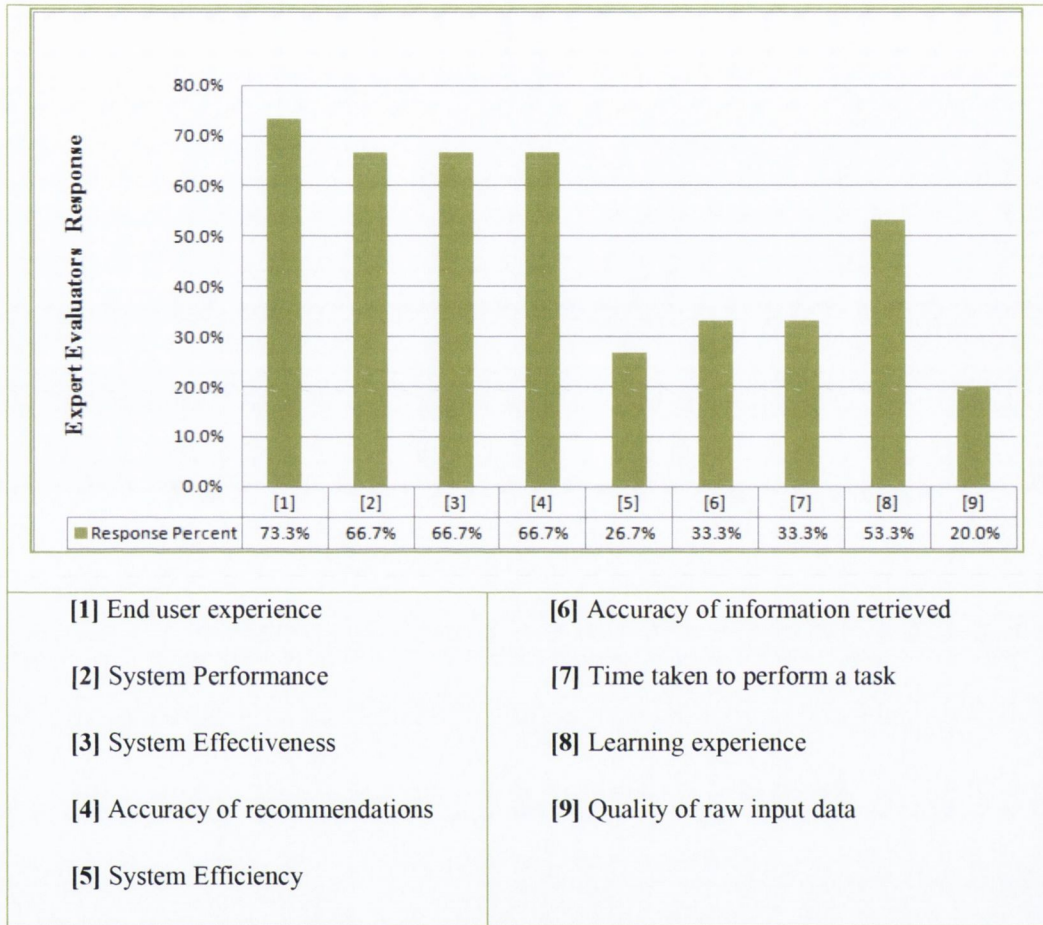
**Figure 6- 13: Expert response on evaluation purpose**

When asked “What kind of question(s) would you wish to answer during evaluation?” 71.8% answered “what is improved by having adaptivity”. Figure 6-15 depicts the response percentage of the 39 experts who responded to this question.



**Figure 6- 14: Response percentage on questions**

In addition, when asked “If you were to conduct the evaluation in the form that you have selected above, what would it help you to improve in the system (i.e. Q3 to Q6)?” 73.3% answered “end-user experience”. A summary of responses is depicted in Figure 6-16.



**Figure 6- 15: Overall response to “what would it help you to improve in the system?”**

In the next step, the experts were asked: “Which of the following evaluation approach(s) would you recommend to be used when evaluating an adaptive system(s) belonging to the variation type (Q2) and properties you identified (Q3 to Q6)? Please rate the recommended approach.” They were also asked to rate the approaches using a Likert scale from 1 (not appropriate) to 5 (most appropriate). Most of the experts rated the user-centred evaluation approach (UCEA) as the most appropriate (25 out of 30), followed by the layered approach (17 out of 30) and the utility-based (10 out of 30). A summary of these results of recommended evaluation approaches rated by experts is presented in Figure 6-17.

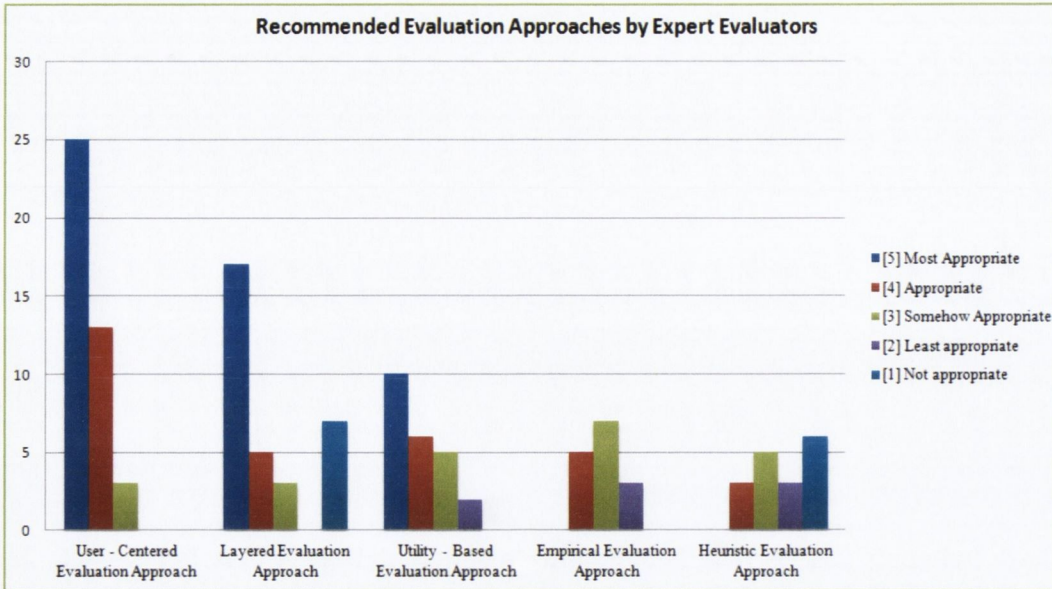


Figure 6- 16: Recommended evaluation approaches

### Appropriate Evaluation Methods

The experts were then asked: “Which of the following evaluation methods do you feel are appropriate for evaluating systems of the variation type you chose in Q2 and the properties (Q3 to Q6)?”<sup>3</sup> They were reminded that they could recommend one or more evaluation methods by rating them from most appropriate to not appropriate. The results consisted of a wide range of diverse of recommendations; for most appropriate evaluation methods, 4 experts stated they would recommend ‘usability testing’, while for appropriate methods; 3 experts would recommend ‘Questionnaires’ respectively. 3 experts agreed somehow appropriate methods would be ‘focus group’. Table 6-1 presents a summary of methods experts would recommend.

**Table 6- 1: Evaluation methods experts would recommend**

Evaluation Methods (Recommend 1)										
Answer Options	Questionnaires	Interviews	Usability Testing	Experimental Evaluation	Focus Group	User Observation	User Test	Expert Review	Quantative	Data Mining
Most Appropriate	1	0	4	0	0	2	2	0	0	2
Appropriate	3	0	0	2	0	2	1	0	0	0
Somehow Appropriate	0	1	0	0	3	2	1	1	0	0
Evaluation Methods (Recommend 2)										
Answer Options	Questionnaires	Interviews	Usability Testing	Experimental Evaluation	Focus Group	User Observation	User Test	Expert Review	Quantative	Data Mining
Most Appropriate	1	0	0	3	0	0	0	0	2	0
Appropriate	0	3	2	0	1	0	0	0	0	1
Somehow Appropriate	0	0	1	0	1	0	1	2	0	0
Evaluation Methods (Recommend 3)										
Answer Options	Questionnaires	Interviews	Usability Testing	Experimental Evaluation	Focus Group	User Observation	User Test	Expert Review	Quantative	Data Mining
Most Appropriate	0	1	0	0	1	1	0	0	0	1
Appropriate	0	1	1	1	0	0	0	0	1	1
Somehow Appropriate	0	1	0	0	1	0	1	1	0	0

Evaluation Methods (Recommend 1)											
Answer Options	Empirical Observations	Simulated Users	Cross-Validation	Heuristic Evaluations	Wizard of Oz Simulation	Creative Brainstorming Sessions	Eye-Tracking	Task Completion Time	System Preference Survey	Formative Evaluation	Summative Evaluation
Most Appropriate	0	0	0	0	0	0	1	0	0	1	0
Appropriate	1	1	0	2	0	0	0	0	0	0	0
Somehow Appropriate	0	0	0	2	0	0	0	0	0	0	0
Evaluation Methods (Recommend 2)											
Answer Options	Empirical Observations	Simulated Users	Cross-Validation	Heuristic Evaluations	Wizard of Oz Simulation	Creative Brainstorming	Eye-Tracking	Task Completion	System Preference	Formative Evaluation	Summative Evaluation
Most Appropriate	0	0	1	0	0	0	1	0	0	0	0
Appropriate	2	1	0	0	0	1	0	0	0	0	0
Somehow Appropriate	0	0	0	0	1	0	0	1	0	0	0
Evaluation Methods (Recommend 3)											
Answer Options	Empirical Observations	Simulated Users	Cross-Validation	Heuristic Evaluations	Wizard of Oz Simulation	Creative Brainstorming	Eye-Tracking	Task Completion	System Preference	Formative Evaluation	Summative Evaluation
Most Appropriate	2	0	0	0	0	0	1	1	0	0	0
Appropriate	0	0	1	1	0	0	0	2	0	0	0
Somehow Appropriate	0	0	0	0	0	1	1	0	0	0	0

***Appropriate Evaluation Criteria***

The next question was: “Which of the following evaluation criteria do you feel are appropriate for evaluating systems of the variation type you chose in Q3?” The experts were reminded that they could recommend one or more evaluation criteria, rating them from most appropriate to not appropriate. The results consisted of a wide range of diverse of recommendations; for most appropriate evaluation criteria, 4 experts would recommend ‘user satisfaction’, while for appropriate criteria; 2 experts would recommend ‘usability’ respectively. 2 experts agreed somehow appropriate criteria would be ‘intention to use’. Table 6-2 presents a summary of evaluation criteria experts would recommend.

**Table 6- 2: Evaluation criteria experts would recommend**

Measurement Criteria (Recommend 1)												
Answer Options	User Satisfaction	User Performance	Usability	Perceived Usefulness	User Skills and Capabilities	Intention to Use	Preferences	Trust and Privacy Issues	Content Adaptation	Appropriateness of Adaptation	User Behaviour	Knowledge of Domain
Most Appropriate	4	2	0	0	0	0	0	0	2	0	0	1
Appropriate	1	0	2	1	0	0	1	1	1	0	1	0
Somehow Appropriate	0	0	1	1	0	2	1	1	1	0	0	0
Measurement Criteria (Recommend 2)												
Answer Options	User Satisfaction	User Performance	Usability	Perceived Usefulness	User Skills and Capabilities	Intention to Use	Preferences	Trust and Privacy Issues	Content Adaptation	Appropriateness of Adaptation	User Behaviour	Knowledge of Domain
Most Appropriate	0	1	0	2	0	0	0	0	1	0	0	0
Appropriate	0	0	1	0	0	1	0	0	0	0	1	0
Somehow Appropriate	0	0	0	0	0	1	0	2	0	0	0	0
Measurement Criteria (Recommend 3)												
Answer Options	User Satisfaction	User Performance	Usability	Perceived Usefulness	User Skills and Capabilities	Intention to Use	Preferences	Trust and Privacy Issues	Content Adaptation	Appropriateness of Adaptation	User Behaviour	Knowledge of Domain
Most Appropriate	2	0	1	2	0	0	0	0	0	0	0	0
Appropriate	0	0	0	0	0	0	0	0	0	1	1	0
Somehow Appropriate	0	0	0	0	0	0	1	0	0	0	1	0
Measurement Criteria (Recommend 1)												
Answer Options	User Goal	Usability of Interface Adaptation	Interface Knowledge	Precision	Recall	Early Prototype Evaluations	Evaluation before Implementation	Transparency	Effectiveness (for decision support)	Appropriateness	User Cognitive Workload	
Most Appropriate	0	0	0	0	0	0	0	0	1	0	0	
Appropriate	0	1	0	0	0	0	0	0	0	0	0	
Somehow Appropriate	1	0	0	0	0	0	0	0	0	0	0	
Measurement Criteria (Recommend 2)												
Answer Options	User Goal	Usability of Interface Adaptation	Interface Knowledge	Precision	Recall	Early Prototype Evaluations	Evaluation before Implementation	Transparency	Effectiveness (for decision support)	Appropriateness	User Cognitive Workload	
Most Appropriate	0	1	0	0	1	0	0	0	0	0	1	
Appropriate	2	0	0	2	0	0	0	0	0	0	0	
Somehow Appropriate	0	1	0	0	0	0	0	0	0	0	1	
Measurement Criteria (Recommend 3)												
Answer Options	User Goal	Usability of Interface Adaptation	Interface Knowledge	Precision	Recall	Early Prototype Evaluations	Evaluation before Implementation	Transparency	Effectiveness (for decision support)	Appropriateness	User Cognitive Workload	
Most Appropriate	1	0	0	0	0	1	0	1	0	0	0	
Appropriate	1	0	0	2	2	1	0	0	0	0	0	
Somehow Appropriate	1	0	1	0	0	0	0	1	1	0	0	

**Appropriate Metrics**

Next, the experts were asked: “Which of the following evaluation metrics do you feel are appropriate for evaluating systems of the variation type you chose in Q3?” and reminded again that they could recommend one or more evaluation metrics (rating them from most appropriate to not appropriate). The results consisted of a wide range of diverse of what metrics the experts would recommend; for most appropriate evaluation metrics, 4 experts would recommend ‘accuracy of recommendations’, while for appropriate evaluation metric. Table 6-3 presents a summary of evaluation criteria experts would recommend.

**Table 6- 3: Experts recommended metrics**

Metrics (Recommend 1)										
Answer Options	Accuracy of Recommendations	Precision	Recall	Accuracy of Retrieval	Reliability Metrics	Task Completion Time	Task Effectiveness	Task Efficiency	Perceived Appropriateness of Adaptations	Invasiveness of Adaptations
Most Appropriate	4	1	0	1	0	1	3	0	0	0
Appropriate	1	0	1	1	0	0	0	1	1	0
Somehow Appropriate	0	1	0	1	1	1	0	1	0	1
Metrics (Recommend 2)										
Answer Options	Accuracy of Recommendations	Precision	Recall	Accuracy of Retrieval	Reliability Metrics	Task Completion Time	Task Effectiveness	Task Efficiency	Perceived Appropriateness of Adaptations	Invasiveness of Adaptations
Most Appropriate	1	1	1	0	2	1	1	0	1	0
Appropriate	0	1	0	1	0	2	0	1	1	0
Somehow Appropriate	0	1	0	0	0	0	2	0	0	2
Metrics (Recommend 3)										
Answer Options	User Satisfaction	User Performance	Usability Accuracy of Recommendations	Precision	Recall	Accuracy of Retrieval	Reliability Metrics	Task Completion Time	Task Effectiveness	Task Efficiency
Most Appropriate	0	0	0	0	1	1	0	0	1	3
Appropriate	1	0	0	0	1	0	0	1	0	1
Somehow Appropriate	2	1	0	0	0	0	0	0	0	0
Metrics (Recommend 1)										
Answer Options	Awareness of Adaptations	Behavioural Complexity	Accuracy of Models	pIA: Performance Influence on Adaptivity	ApOC: Adaptive Personalisation Overall Cost	Students' Content Learning Gains	Presence Questionnaire Scores	Intrinsic Motivation Inventory Scores	Response Count	
Most Appropriate	0	0	1	0	0	0	0	0	11	
Appropriate	1	0	0	0	1	1	0	0	8	
Somehow Appropriate	0	1	1	0	0	0	0	0	8	
Metrics (Recommend 2)										
Answer Options	Awareness of Adaptations	Behavioural Complexity	Accuracy of Models	pIA: Performance Influence on Adaptivity	ApOC: Adaptive Personalisation Overall Cost	Students' Content Learning Gains	Presence Questionnaire Scores	Intrinsic Motivation Inventory Scores	Response Count	
Most Appropriate	0	0	0	0	0	0	0	0	8	
Appropriate	1	1	0	0	0	0	0	0	8	
Somehow Appropriate	1	0	0	0	0	0	0	0	6	
Metrics (Recommend 3)										
Answer Options	Perceived Appropriateness of Adaptations	Invasiveness of Adaptations	Awareness of Adaptations	Behavioural Complexity	Accuracy of Models	pIA: Performance Influence on Adaptivity	ApOC: Adaptive Personalisation Overall Cost	Students' Content Learning Gains	Presence Questionnaire Scores	
Most Appropriate	0	0	0	0	0	0	0	0	0	
Appropriate	0	1	0	0	0	2	0	0	0	
Somehow Appropriate	0	0	0	1	0	0	0	0	1	

**Appropriate Bundle**

Furthermore, the experts were asked: “Which of the following evaluation techniques (i.e. the techniques you recommended in Q9-11) would you bundle to be used together? The term “bundle” refers to appropriate combination of a method/criteria/metric that can be used together when evaluating the properties you selected in Q4-Q7?” The results consisted of a wide range of diverse of what bundles (method, metric and criteria) the



experts would recommend to be used together; 3 experts agreed most appropriate bundle was ‘*bundle 1*’ for evaluation methods. In bundling section, 4 experts agreed the most appropriate bundle was ‘*bundle 1*’ for evaluation metrics 3 experts also agreed on ‘*bundle 1*’. A summary of bundles experts would recommend is presented in Table 6-4.

**Table 6- 4: Expert -Response recommended bundles (method, criteria and metric)**

Evaluation Methods													
Answer Options	Questionnaires	Interviews	Usability Testing	Experimental Evaluation	Focus Group	User Observation	User Test	Expert Review	Quantative	Data Mining	Empirical Observations	Simulated Users	Cross-Validation
Most Appropriate Bundle (Bundle 1)	3	0	1	0	0	3	2	0	0	1	1	0	0
Appropriate Bundle (Bundle 2)	0	1	2	1	1	0	1	0	0	0	1	0	1
Somehow Appropriate (Bundle 3)	0	1	0	1	0	0	1	0	2	0	0	0	0
Measurement Criteria													
Answer Options	User Satisfaction	User Performance	Usability	Perceived Usefulness	User Skills and Capabilities	Intention to Use	Preferences	Trust and Privacy Issues	Content Adaptation	Appropriateness of Adaptation	User Behaviour	Knowledge of Domain	User Goal
Most Appropriate Bundle (Bundle 1)	4	3	0	0	0	0	0	0	0	0	0	1	0
Appropriate Bundle (Bundle 2)	1	1	1	1	0	0	0	0	0	1	0	0	0
Somehow Appropriate (Bundle 3)	1	2	0	1	0	0	0	0	0	0	0	0	0
Evaluation Metric													
Answer Options	Accuracy of Recommendations	Precision	Accuracy of Retrieval	Reliability Metrics	Task Completion Time	Task effectiveness	Task efficiency	Perceived appropriateness of adaptations	Invasiveness of adaptations	Awareness of adaptations	Behavioural Complexity	Accuracy of models	pIA: Performance Influence on Adaptivity
Most Appropriate Bundle (Bundle 1)	3	1	0	0	0	1	0	3	1	0	1	0	0
Appropriate Bundle (Bundle 2)	1	0	0	0	0	2	0	0	1	2	0	1	0
Somehow Appropriate (Bundle 3)	0	1	0	0	1	1	1	1	0	0	0	0	0

In conclusion the results by experts consisted of a wide range of diverse of what evaluation techniques and bundles they would recommend. In an ideal situation the candidate would be able to get the experts to describe how they would evaluate individual systems; however when we interviewed them they said they won’t have hands on experience on recommending evaluation techniques for systems developed by other people they would be able to recommend techniques based on generic systems belong to certain variation types (e.g. adaptive educational hypermedia systems).

**6.2.2.4 Comparison-: Expert Results vs Recommendations produced by the Recommender System**

If this is what the experts are saying (section 6.2.2), now how do that compare with the hybrid recommender system. Well the system works on specific systems rather than generic systems. Furthermore in most cases adaptive systems evaluators won’t have

access to an expert, however what we have been able to show is that the recommender systems seem to align with what the experts are recommending (see table 6-5).

**Table 6- 5: Results (evaluation techniques) produced by the hybrid recommender system**

System Name	Recommended Evaluation Techniques		
	Methods	Criteria	Metrics
APeLs	Task-based , usability testing, Expert Review, Pre and post questionnaires	Usability, Perceived Usefulness, Appropriateness of Adaptation	Perceived appropriateness of adaptations, Invasiveness of adaptations, Awareness of adaptations
ARCHING	Interviews, Questionnaires, User Observation, Usability Testing, Data Mining, Simulated Users, Cross-Validation	Usability, Perceived Usefulness, Intention to Use, Appropriateness of Adaptation, User Satisfaction, Content Adaptation, User Performance, Real User Actions	Appropriateness of Adaptation, User satisfaction
PEACH	Interviews, Questionnaires, ask-based, Quantative methods , Simulated Users, Experimental Evaluation, Empirical Observations	Usability, Perceived Usefulness, Intention to Use, User Behaviour, User Satisfaction, Early Prototype Evaluations	Accuracy of Recommendations, User satisfaction, use of recommended items
ERM-Tutor (2005)	Questionnaires, Focus Group, User Observation, Expert Review, Wizard of Oz Simulation, Usability Testing, Experimental Evaluation, Data Mining	Usability, Perceived Usefulness, User Behaviour, Content Adaptation, Preferences, User Performance, User Cognitive Workload, Real User Actions	User satisfaction, Performance

Furthermore, when the candidate sort to gain more insight into which features of the hybrid recommender system (provision of explanations regarding recommended techniques) the experts would find useful (section 6.2.2.3, Figure 6-11) i.e.:

- Explanations of recommended evaluation approach (es)
- Explanations of recommended evaluation method (s)
- Explanations of recommended measurement criteria (s)
- Explanations of recommended measurement metric (s)
- Explanations of recommended bundles (method, criteria and metric)

When the candidate compared the results (provision of explanations) recommended by the experts (Figure 6-11), to those results produced by the recommender system, they were the same. In this case we can argue that the hybrid system is producing good results.

## **6.3 Personalised Search System: Search Identification, User Satisfaction and Learnability**

### **6.3.1 Experiment Objective**

The benefit to the evaluator who needs evaluation studies of adaptive systems lies in a system's ability to assist a user's search for information effectively and efficiently. In particular, it is desirable that a system requires users to invest the least amount of effort in finding relevant information as quickly as possible.

The evaluation objectives for this experiment were as follows:

**Evaluation Objective 1:** The first evaluation objective of the novice evaluators regarding user efficiency and effectiveness and evaluates how well; The Personalised Search System supports novice evaluators during *search identification* of relevant evaluation studies of adaptive systems. In addition the system should enable the users to *browse, view and retrieve* relevant search results for queries on evaluation of adaptive systems. Furthermore the *presentation of the search results*, need to be 'helpful'.

**Evaluation Objective 2:** In terms of user satisfaction, the benefit to evaluators lies in the perceived usability of the various functionalities provided by the personalised search system. In particular, the assumption is that evaluators recognise and value the personalised search system's various functionalities. The aim of the second objective was to find out whether users were satisfied after interacting with the system. In addition evaluators recognise and value the presentation of returned results (i.e. finding evaluation studies of internal models of adaptive systems, evaluation studies of adaptive systems and general evaluation studies of such systems). Furthermore the experiment was also focused on finding out if evaluators would *not need the support* of a technical person to be able to use the personalised search system. This would help us know whether users were able to learn more about evaluations of adaptive systems. Learnability is used to describe the

ability of the personalised search system interface to allow users to accomplish tasks at the first attempt.

In order to test this objective, usability scores (of Q4 and Q10 of SUS questionnaire) are used.

### 6.3.2 Experimental Setup

A total of 45 users participated in this experiment; of these, 43 completed the full evaluation process. Participants were recruited from Trinity College Dublin, Dublin Institute of Technology, the UMAP (2011, 2012 and 2013) conference and the AH conference, DataTEL and Recommender Communities. The experiment aimed at identifying user appreciation and satisfaction regarding the various functionalities provided by the personalised search system.

The users were informed in this task that they would be interacting with a personalised search system that searched across a knowledge repository of an educational evaluation dataset extracted from over 450 studies of adaptive systems, published from 2000 to 2012. The system has three main features that allow novice and expert evaluators to search for:

- i) Evaluation Studies of Internal models of Adaptive Systems
- ii) Evaluation Studies of Adaptive Systems.
- iii) General Evaluation Studies of Adaptive Systems

Figure 6-18 depicts the experimental setup process.

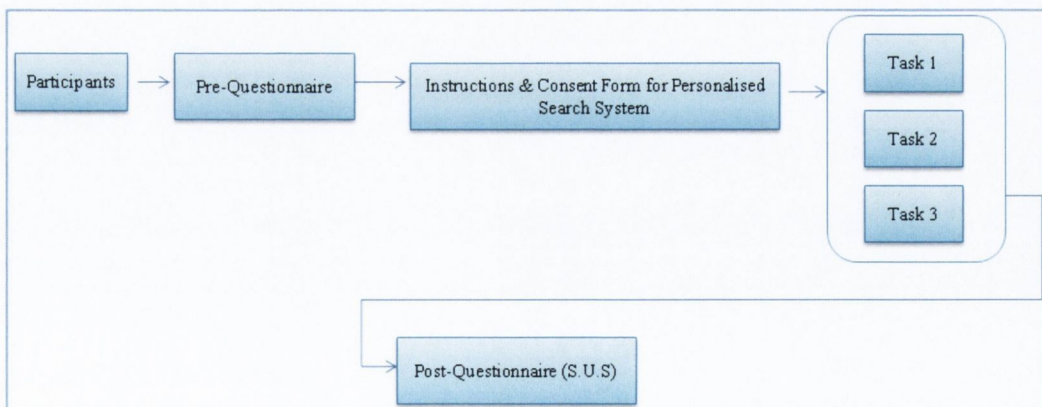
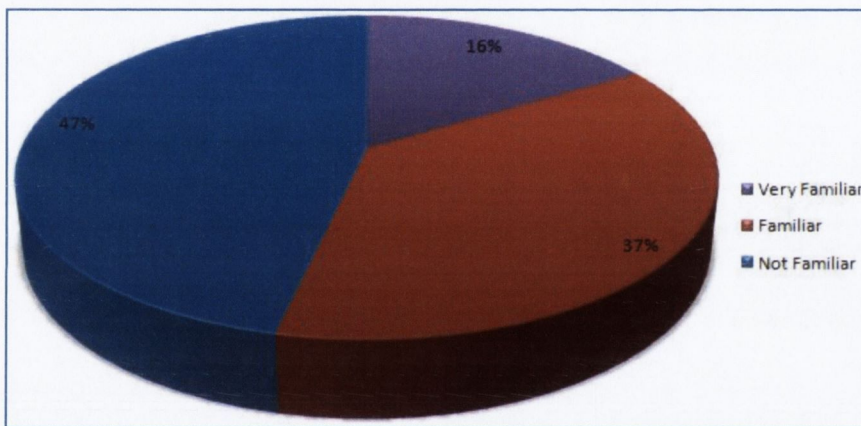


Figure 6- 17: Experimental setup of the personalised search system

### 6.3.3 Results and Findings

#### *User characteristics*

The results revealed that the majority of the users were novices. The user characteristics gathered from the pre-questionnaire are presented in Figure 6-19. As can be seen from this figure, most participants had not used personalised search systems. When asked “*How familiar are you with personalised search systems (PIS) that allow to you find evaluation studies of adaptive systems?*” 47% stated *not familiar*, while 16% stated *very familiar*. These results are significant because the personalised search system is aimed at supporting novice evaluators of adaptive systems. A summary of all the characteristics is depicted in Figure 6-19.



**Figure 6- 18: User characteristics of personalised search system participants**

In order to identify how experienced the participants were and how frequently they used personalised search systems to find evaluation studies of adaptive systems in general, the first set of questions required evaluators to select from a Likert scale of 1 (never), 2 (once or twice), 3 (sometimes), 4 (regularly) and 5 (very often). Participants were asked “*How often do you use personalised search system to find studies which detail evaluations of adaptive systems?*” 37% had *never* used such a system, 23% had used it *once or twice* and 27% stated *sometimes*. A summary of the users’ responses is presented in Figure 6-20.

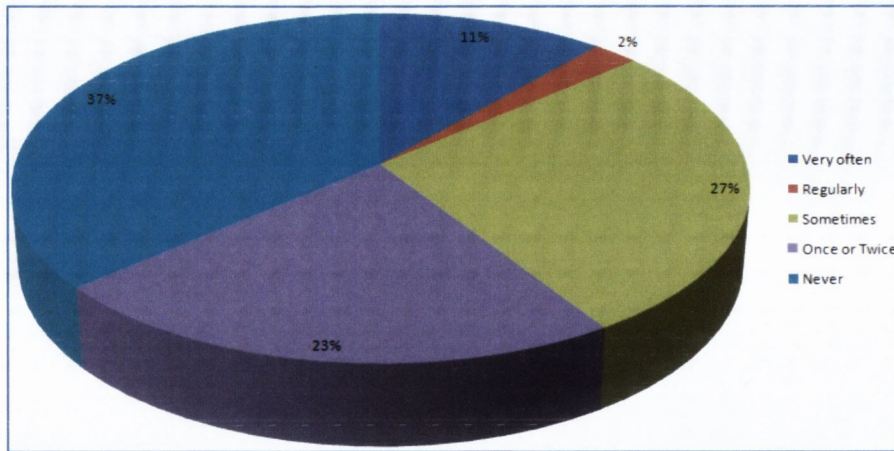


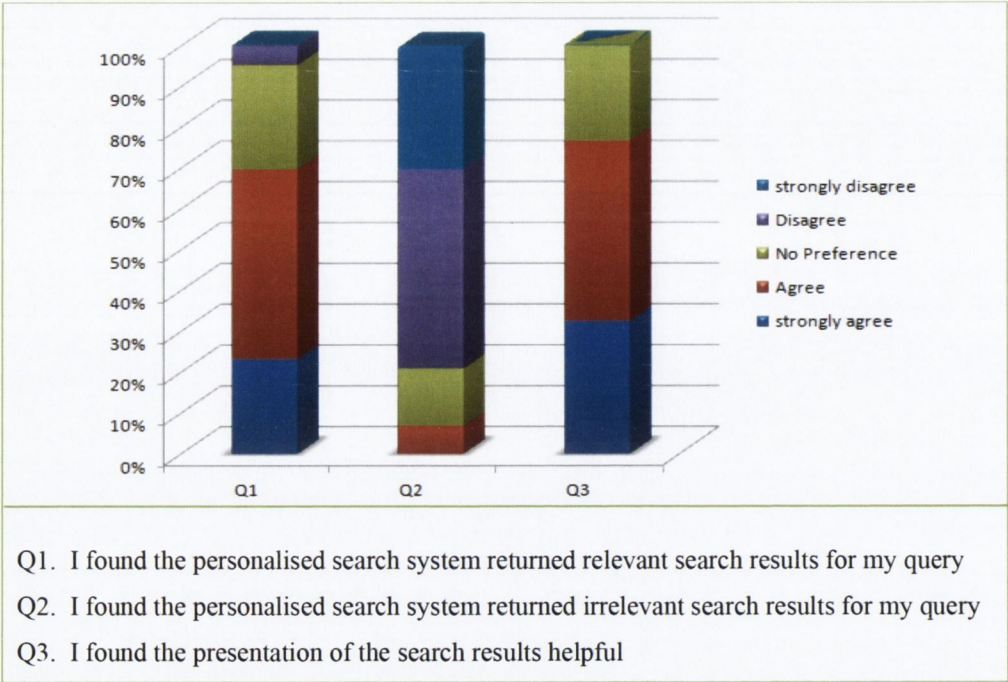
Figure 6- 19: Familiarity in using personalised search system

### *Task assistance*

As stated in the experimental set-up section (section 6.3.2), participants were asked to complete a set of usability questionnaires after using the personalised search system to search for: (i) evaluation studies of internal models of adaptive systems (Manouselis et al.), (ii) evaluation studies of adaptive systems and (iii) general evaluation studies of adaptive systems. The first set of questions asked users to select from statements on a Likert scale of 1 (strongly disagree) to 5 (strongly agree), including also the Standard Usability Scale (SUS). A second set of questions allowed users to express freely any particular likes and dislikes concerning the personalised search system.

Participants were presented with three tasks. Task 1 involved finding: (i) evaluation studies of internal models of adaptive systems. When asked to agree or disagree with the statement “*I found the personalised search system returned relevant search results for my query*”, 69.76% agreed, while 76.74% disagreed with the statement “*I found the personalised search system returned irrelevant search results for my query*”.

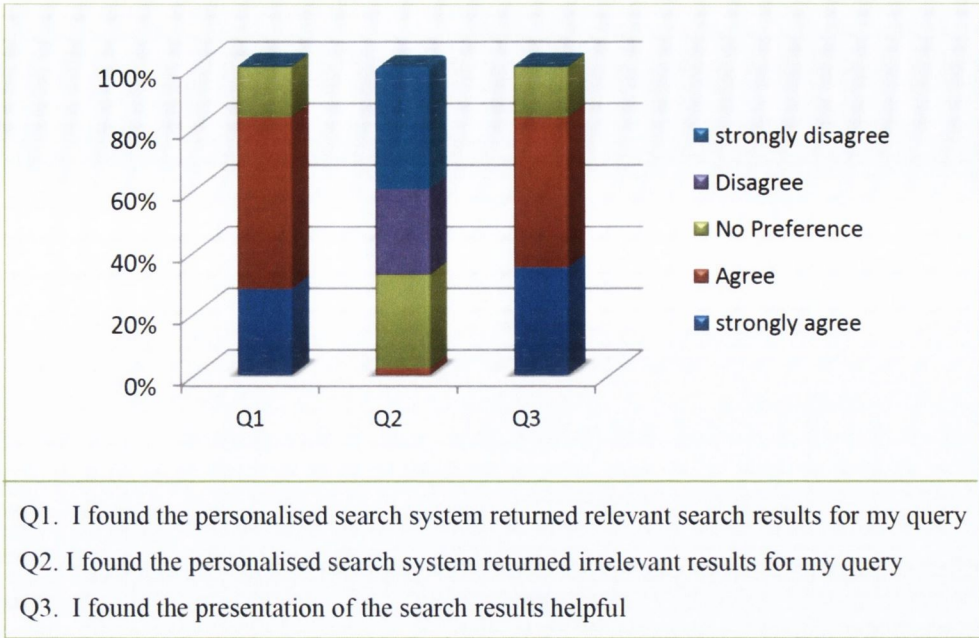
When asked to agree or disagree with the statement “*I found the presentation of the search results helpful*”, 77% agreed. The results for task 1 are depicted in Figure 6-21.



**Figure 6- 20: Responses on return of relevant search results of internal models**

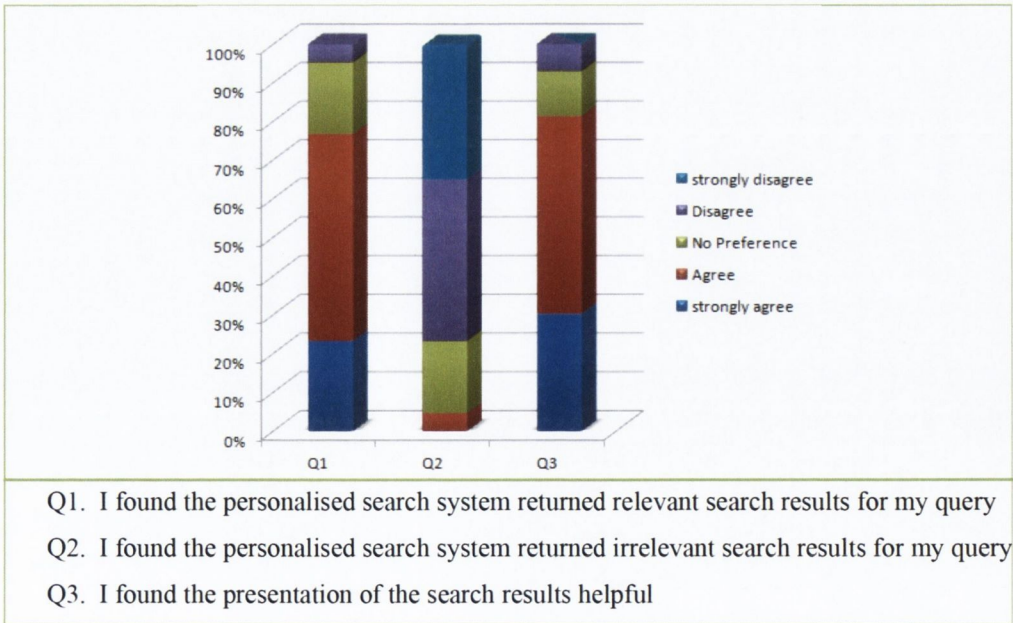
The second task involved participants interacting with the personalised search system to find evaluation studies of adaptive systems. When asked to agree or disagree with “I found the personalised search system returned relevant search results for my query”, 82.73% agreed, while 67.4% disagreed.

The participants were further asked to agree or disagree with the statement “I found the presentation of the search results helpful”; 83.72% agreed. A similar percentage disagreed with the statement “I found the personalised search system returned irrelevant results for my query”. The results for task 2 are depicted in Figure 6-22.



**Figure 6- 21: Responses on return of relevant search results of evaluations of adaptive systems**

In the third task, participants were required to interact with the search system and find general evaluation studies of adaptive systems. When asked to agree or disagree with the statement “*I found the personalised search system returned relevant search results for my query*”, 76.74% agreed. When asked to agree or disagree with the statement “*I found the presentation of the search results helpful*”, 81.39% agreed. The results for task 3 are depicted in Figure 6-23.



**Figure 6- 22: Task 3 – Responses on return of relevant search results of general evaluation studies**

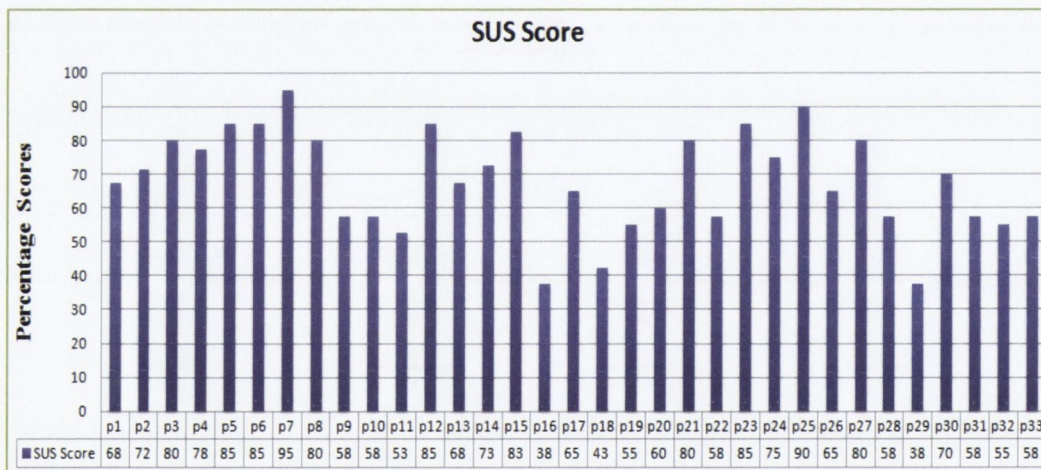


## Usability

In addition to the task-based questions, the user study aimed at identifying users' appreciation and satisfaction regarding the various functionalities provided by the personalised search system. Out of the 43 users who participated in this study, 33 completed the SUS questionnaire.

As mentioned in the experimental setup section (6.3.2), users were also asked to complete a set of usability questionnaires after searching for information on evaluations of adaptive systems. The second set of questions required users to agree with a set of Likert scale items from 1 (strongly disagree), 2 (disagree), 3 (no preference), 4 (agree) and 5 (strongly agree), using the Standard Usability Scale (SUS).

First of all, in order to determine the overall usability, SUS scores were calculated. The personalised search system scored an average of 68.0; these results are depicted in Figure 6-24.



**Figure 6- 23: A summary of percentile scores of 33 participants of the personalised search system**

### Usability (User satisfaction)

When asked to agree or disagree with the statement *“I think that I would like to use this Personalised Search System frequently”*, 48% agreed while a small percentage (6%)

disagreed. When asked to agree or disagree with the statement *"I thought the Personalised Search System was easy to use"*, 45% agreed and 6% disagreed.

Furthermore when asked to agree or disagree with the statement *"I found the various functions in this Personalised Search System were well integrated"*, 48% agreed while 3% disagreed. Asked to agree or disagree with the statement *"I would imagine that most people would learn to use this Personalised Search System very quickly"*, 30%/36% strongly agreed/agreed while 6% disagreed. Asked to agree or disagree with the statement *"I felt very confident using the Personalised Search System"*, 24%/52% strongly agreed/agreed 6% disagreed.

Participants also strongly disagreed/disagreed that the personalised search system was unnecessarily complex, inconsistent or cumbersome to use. The results are outlined below.

When asked to agree or disagree with the statement *"I found the Personalised Search System unnecessarily complex"*, 18%/39% strongly disagreed/disagreed while 9% agreed. Asked to agree or disagree with *"I thought there was too much inconsistency in this Personalised Search System"*, the majority of evaluators strongly disagreed/disagreed (27%/39%) while 3% agreed. Finally, asked to agree or disagree with the statement *"I found the Personalised Search System very cumbersome to use"*, most evaluators strongly disagreed/disagreed (15%/36%) while 18% agreed. A summary of these results is presented in Figure 6-25

### ***Usability (Learnability)***

Questions 4 and 10 of SUS focus on the learnability dimension. The results of user's responses to these two questions are depicted in Figure 6-25. When asked to agree or disagree with the statement *"I think that I would need the support of a technical person to be able to use this Personalised Search System"*, the majority of the evaluators strongly disagreed/disagreed (39%/36%) while 21% agreed. Asked to agree or disagree with the statement *"I needed to learn a lot of things before I could get going with this Personalised Search System"*, most evaluators strongly disagreed/disagreed (30%/27%) while 3% strongly agreed.

Overall, these are very encouraging results (user satisfaction and learnability) for such a system, especially considering that the majority of the users had not used a personalised search system for evaluation studies of adaptive systems before. A summary of these results is depicted in Figure 6-25.

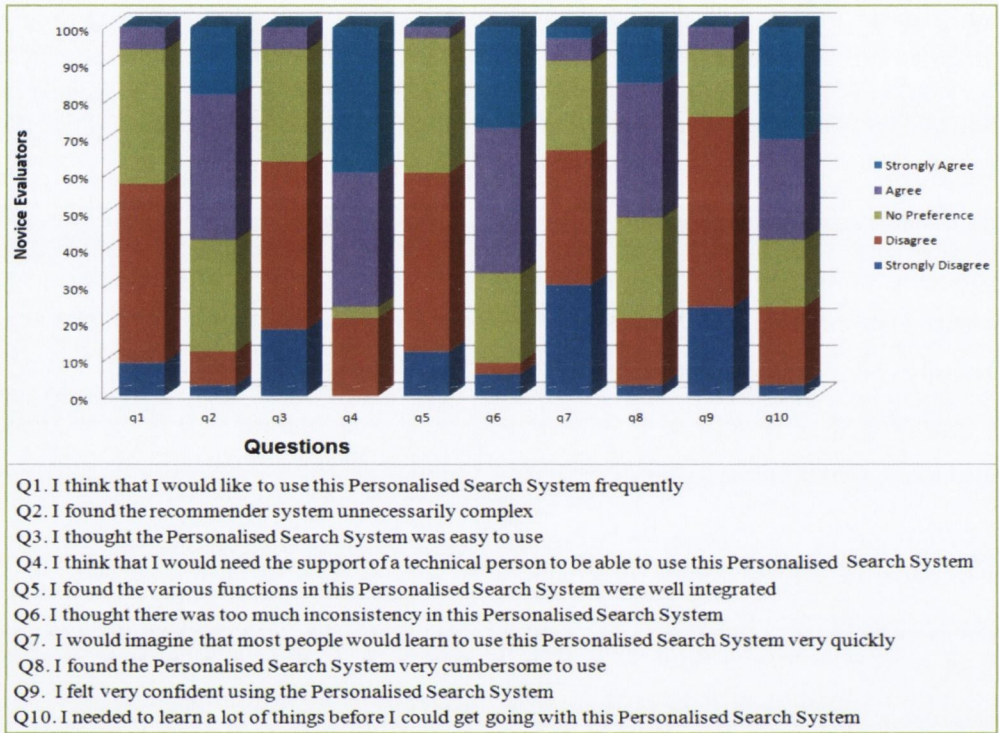


Figure 6- 24: User satisfaction

In addition, due to the large number of data points, I also computed the frequency of distribution, in order to visualize variability. These results are presented in Figure 6-26.

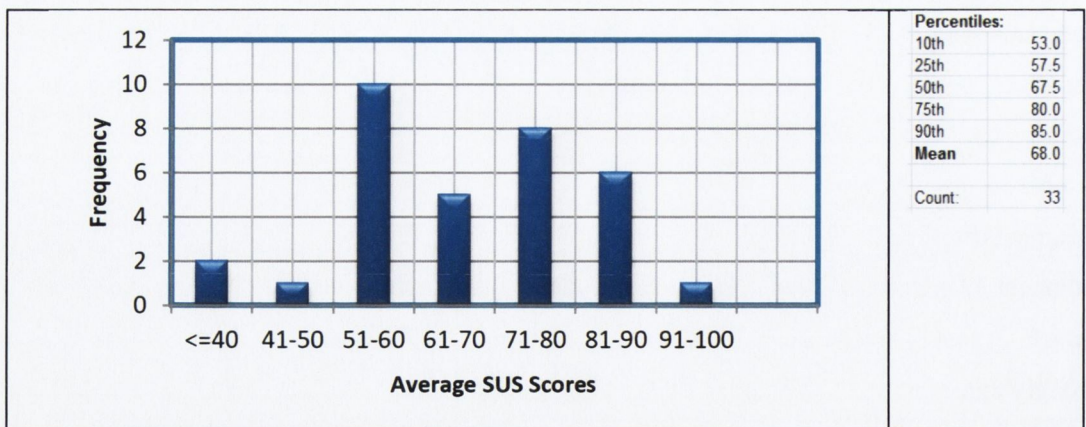


Figure 6- 25: Frequency distribution of 33 participants – personalised search system

In conclusion, based on the results and findings of the novice evaluators after completing experiment 2 (research objective 4), the evaluations of the novices seem to prove that both the evaluations made by the personalised search sub-system seem credible, well argued and well backed up and therefore majority of them said that they were able to identify evaluation studies of adaptive system.

## **6.4 Taxonomy of Technical Terms: Usability**

### **6.4.1 Experiment Objective**

This user study aimed at identifying users' appreciation and satisfaction regarding the various functionalities provided by the taxonomy of technical terms of evaluation of adaptive systems.

In terms of user satisfaction, the benefit to evaluators lies in the perceived usability of the various functionalities provided by the taxonomy of technical terms for evaluations of adaptive systems. In particular, the assumption is that evaluators recognise and value the taxonomy's various functionalities. This objective aimed at finding out whether users satisfied after using the taxonomy. Furthermore users understand (learn) and value the presentation of the technical terms. The experiment also aimed at finding out if evaluators would *not need the support* of a technical person to be able to use the taxonomy.

### **6.4.2 Experiment Setup**

A total of the 18 novice evaluators participated in this experiment, out of which 15 completed the experiment. Participants were recruited from Trinity College Dublin, Dublin Institute of Technology, the UMAP (2011, 2012 and 2013) conference and the AH conference, DataTEL and Recommender Communities. Evaluators were asked after interacting with the taxonomy of technical terms to complete a SUS questionnaire. The set of questions asked required evaluators to agree with Likert scale items from 1 (strongly disagree), 2 (disagree), 3 (no preference) to 4 (agree) and 5 (strongly agree), using the SUS.

### 6.4.3 Results and Findings (SUS Scores)

In order to determine the overall usability, SUS scores were calculated. The taxonomy of technical terms scored an average of 86.3%; these results are depicted in Figure 6-27.

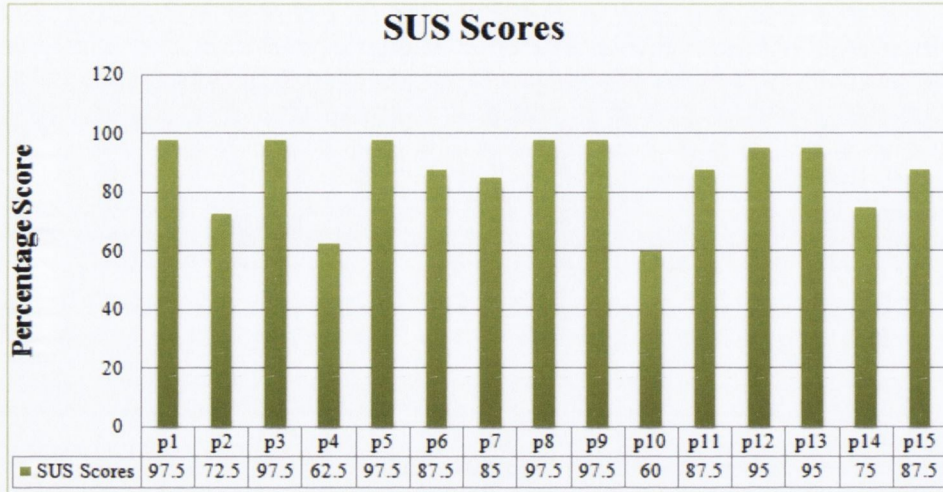


Figure 6- 26: Summary of percentile scores of the taxonomy

#### *User satisfaction*

When asked to agree or disagree with the statement *“I think that I would like to use this taxonomy”*, 53% agreed that they would use the taxonomy frequently. Asked to agree or disagree with the statement *“I thought the taxonomy was easy to use”*, the majority of the evaluator’s strongly agreed/agreed (47%/40%) the taxonomy was easy to use and 13% disagreed.

Furthermore, when asked to agree or disagree with the statement *“I found the various functions in this taxonomy were well integrated”*, 43%/33% strongly agreed/agreed while 20% had no preference. Asked to agree or disagree with the statement *“I would imagine that most people would learn to use this taxonomy very quickly”*, 47%/33% strongly/agreed while 20% disagreed. Asked to agree or disagree with the statement *“I felt very confident using the taxonomy”*, 47%/33% strongly agreed/agreed while 20% had no preference.

Participants also strongly disagreed/disagreed that the taxonomy was unnecessarily complex, had inconsistencies and was cumbersome to use. The results are shown below.



Figure 6- 27: User satisfaction and learnability

When asked to agree or disagree with the statement “I found the taxonomy unnecessarily complex”, 47%/47% strongly disagreed/disagreed while 7% had no preference. Asked to agree or disagree with the statement “I thought there was too much inconsistency in this taxonomy”, the majority of evaluators strongly disagreed/disagreed (80%/20%). Finally, when asked to agree or disagree with the statement “I found the taxonomy very cumbersome to use”, the majority of evaluators strongly disagreed/disagreed (87%/13%). A summary of these results is presented in Figure 6-28.

### Usability (Learnability)

Questions 4 and 10 of SUS provide the learnability dimension. The results of users responding to these two questions are depicted in Figure 6-30. When asked to agree or disagree with the statement “I think that I would need the support of a technical person to be able to use this taxonomy”, the majority of the evaluators strongly disagreed/disagreed (80%/20%). Asked to agree or disagree with the statement “I needed to learn a lot of

*things before I could get going with this taxonomy*", most evaluators disagreed (80%) while 7% agreed.

Overall, again, these are very encouraging results concerning user satisfaction and learnability for such a taxonomy of technical terms for evaluation of adaptive systems, especially considering that the majority of the users had not used such taxonomy before.

Due to the large number of data points, I also computed the frequency of distribution, in order to visualize variability. In conclusion, based on the results and findings of the novice evaluators after completing experiment (research objective 4), the evaluations of the novices seem to prove that both the evaluation results produced by the taxonomy seem credible, well argued and well backed.

## **6.5 Conclusions**

The evaluation results have revealed the benefits of the hybrid recommender system approach to recommending accurate and appropriate evaluation approaches and techniques (methods, metric and criteria) as it has been shown to significantly enhance evaluators' satisfaction and learnability. Similar to the findings in Chapter 4, it is shown that the hybrid recommendation approach supports, encourages and motivates users to learn more about evaluation of adaptive systems. Compared to human experts, the recommender system performs better and produces more accurate and appropriate results. However it is difficult to make strong claim that the hybrid recommender system is as good as what would be recommended by the human experts. One of the reasons is because it is difficult to get enough expert evaluators and also when conducting the experiment, we found it difficult to make complete comparable tests.

The provision of explanations on how such recommended evaluation approaches and techniques are derived is significant (discussed in sections 4.3.1.1 and 4.4.1)

Furthermore, the evaluation results have revealed the educational benefits of the personalised search system for novice evaluators when it comes to task assistance, user satisfaction, effectiveness of results and overall learnability. Users were satisfied with the overall performance of the system. In addition, the results of the taxonomy of technical terms are encouraging; evaluators were satisfied with its performance.

Overall, this chapter has revealed that both novice and expert evaluators acknowledge that all the three components of EFEx framework are valuable and appropriate, especially in supporting the evaluation of adaptive systems, and specifically adaptive TEL systems. These findings are very encouraging and similar to the findings in Chapter 4, section 4.4.



# **Chapter 7: Focused Online Crawling Systems for Evaluation Studies**

## **7.1 Introduction**

Section 7.3.1 presents a focused online crawling system, organized in a way that supports reasoning about the structures of the system. The system architecture comprises three major components: RSS feed management, RSS feed crawler and published study crawling management module.

Chakrabarti et al. (1999) define a focused crawler as a “web crawler which actively seeks, acquires indexes and maintains pages on a specific topic which represent a relatively narrow segment of the WWW” (Chakrabarti et al., 1999). Besides sourcing content based on its content, focused crawling allows a web crawler to process specific sites to greater depths than general-purpose crawlers. Furthermore, focused crawlers can spend more time perusing highly relevant sites rather than attempting to attain broad coverage of the entire WWW in a breadth-first manner. As a result, highly relevant pages can be discovered that may have been overlooked by more general-purpose crawlers (Chakrabarti et al. 99).

This section briefly presents an overview of the influences from the literature review described in Chapter 2. Based on these influences, the architecture, implementation and evaluation of a focused online crawling system for evaluation studies of adaptive system are discussed. Section 7.3 discusses the system architecture and technological design. Section 7.4 describes the implementation of these components in two parts: the administrator interfaces and the novice evaluator interface. Section 7.5 presents the evaluation results and finally section 7.6 concludes the chapter.

## **7.2 Objectives and Scope of OSSES**

The main goal of the focused web crawling system was to selectively seek out pages that are relevant to a pre-defined set of topics in the area of evaluations. The system should be capable of crawling evaluation studies published online in pdf and store them in a centralized repository for further processing.

## 7.3 Architecture and Technical Design

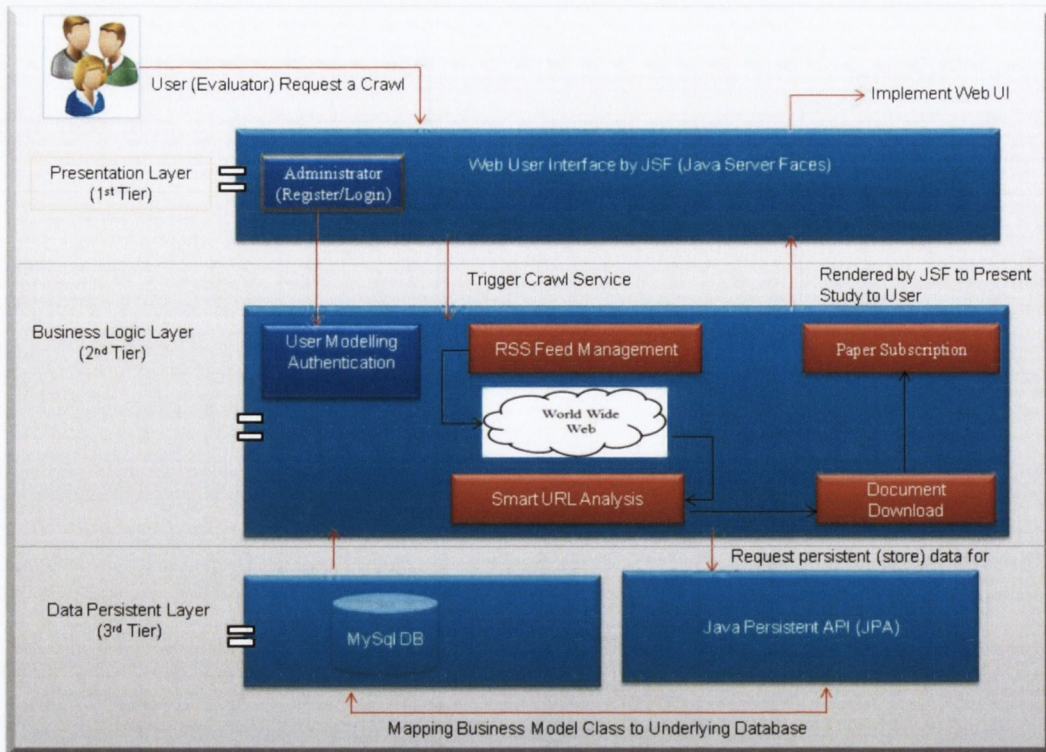
To identify theoretically needs mapping; there is a need to develop a focused crawling system for evaluation studies of adaptive systems. The use of a web crawler is useful for reducing the time taken to complete the task of searching for published evaluation studies, especially by researchers who are new in the area of evaluations of adaptivity. Furthermore, due to the rapid increase of web-linked information on the WWW, it has become difficult for search engines to find exactly appropriate information. The use of large-scale search engines such as Google is very common in surfing the World Wide Web. The capability of these search engines is impressive. Search engines have five components: a crawling module, an indexing module, a page ranking module, a search module, and a page repository (Olston and Pandey, 2008). The crawling module is responsible for the process of downloading documents from the Internet. This process is done by web crawlers which start with a set of seed URLs, and download web pages and extra links from the downloaded pages for further download. The behaviour of a web crawler<sup>34</sup> is the outcome of a combination of policies (Girardi et al., 2006).

### 7.3.1 Architectural Design

The term architecture here refers to the conceptual model that defines the structure, behaviour and views of the focused crawling system (Jaakkola H. and B., 2011). This section describes the proposed crawling system architecture (Figure 7-1).

---

<sup>34</sup> A web crawler is a program that automatically traverses the web's hyperlink structure and retrieves information for the user.

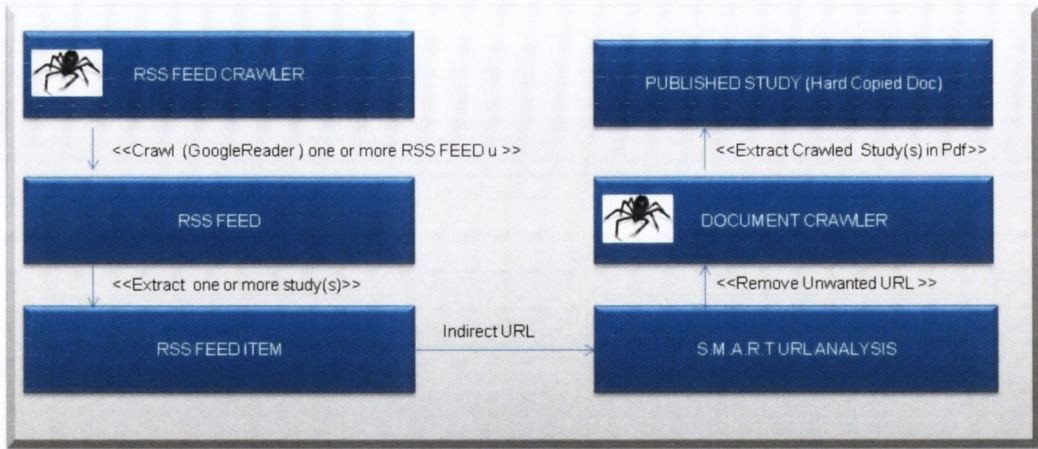


**Figure 7- 1: High-level overview of various components of the focused crawling system**

The crawling process (Figure 7-2) is triggered by the system administrator (novice or expert evaluator). As the evaluator triggers the crawling process, an RSS (Rich Site Summary) web feed<sup>35</sup> (GoogleReader), which includes summarized text, plus metadata such as publishing dates and authorship, is retrieved by the RSS feed crawler. Next, the RSS feed crawler sends a request to get the most recently published papers, and then it automatically creates one or more RSS feed items. The RSS feed item contains the meta-data about the published papers, such as the title, author, published date-time and a URL to the paper document. Subsequently, the URL to the paper document is passed through the Self-Monitoring, Analysis, and Reporting Technology, S.MAR.T URL ANALYSIS. If the URL is a downloadable document link, the analyzer will leave the URL untouched. Otherwise, the analyzer will try to ascertain the downloadable URL for the paper, as discussed above. Finally, the downloadable URL is passed to the Document crawler, which uses the client URL (cURL<sup>36</sup>) to retrieve the document and create a hard pdf copy in the local document repository.

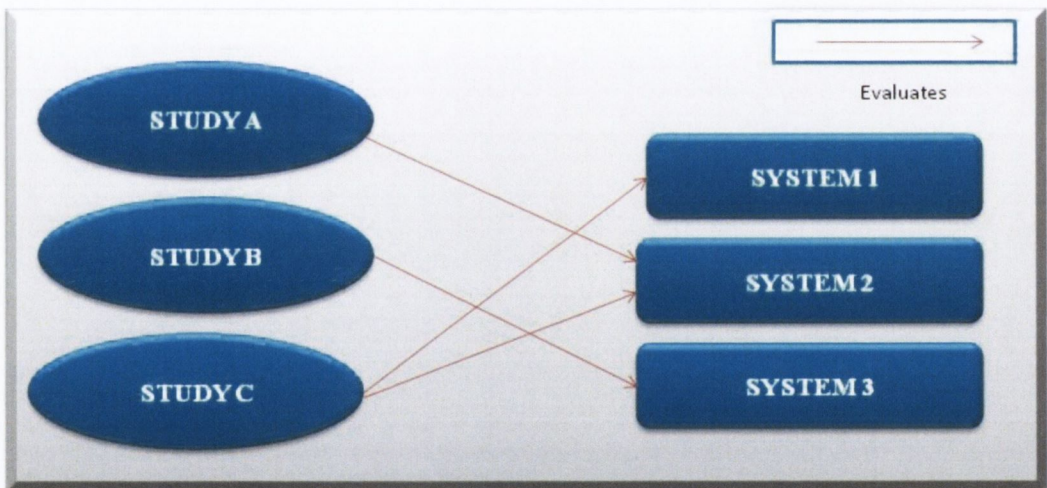
<sup>35</sup> "Web feeds | RSS | The Guardian | guardian.co.uk", *The Guardian*, London, 2008, webpage: GuardianUK-webfeeds.

<sup>36</sup> A command line tool for getting or sending files using URL syntax



**Figure 7- 2: Process of crawling a published study**

The relationship between an evaluation study and an evaluated system is demonstrated in Figure 7-3 for example, while one study might contain the evaluation results and findings of one or more systems (e.g. Study C), a system might be evaluated and the results published in one or more studies (e.g. System 2).

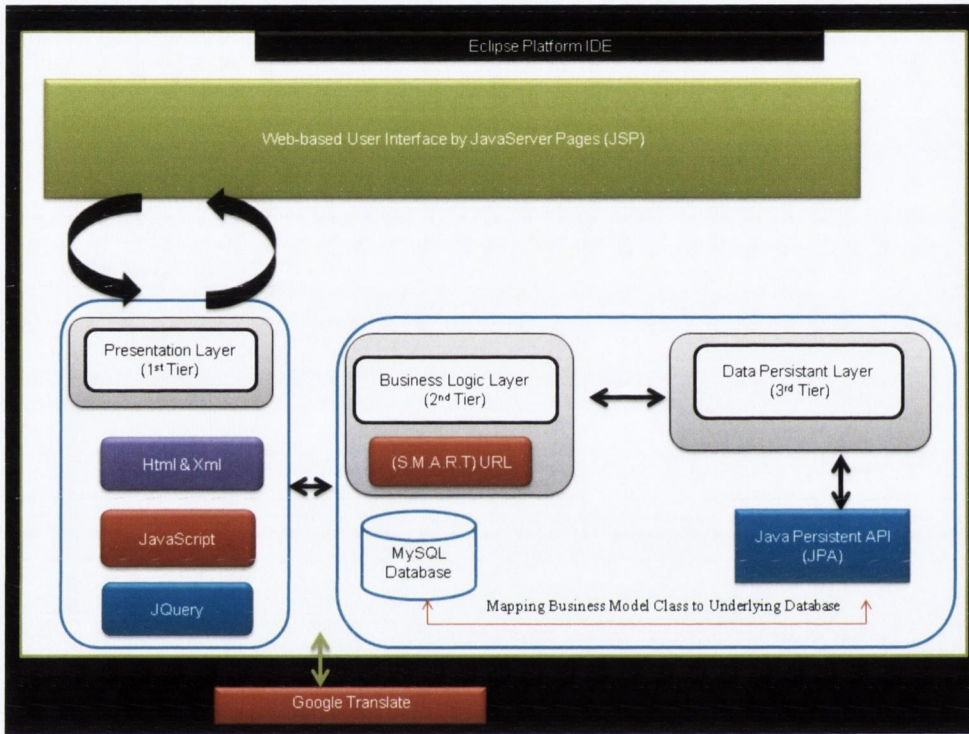


**Figure 7- 3: Relationship between an evaluation study and an evaluated system**

### 7.3.2 Technical Design

The different components of the crawling system are implemented by integrating several technologies and software programs (Figure 7-4). In the third tier (data persistent layer) JPA is used to insert records into the database. The JPA takes the business model ‘study’ and persists; it into the mySql ‘Estudy’ database. In the second tier (business logic layer),

the RSS Feed Crawler crawls the RSS Feed and extracts the study from the RSS Feed. Finally, in the first tier (presentation layer) the study is rendered by JSF to present the Web User Interface to the user. The Eclipse platform was used to develop the system. The platform was chosen because it defines a set of frameworks and common services that collectively make up the ‘integration ware’ required to support a comprehensive tool integration platform.



**Figure 7- 4: Technical design of OSSES system**

Furthermore, the platform defines a workbench user interface and a set of common domain-independent user interaction paradigms that enable plugging into and adding new capabilities to the system. Apache-Openjpa was used to store and retrieve data from the database. The candidate used two servers (Apache Tomcat Server and MySql database server). For Java database connectivity, MySql connector-java was used. To parse the RSS feed, JSON (JavaScript Object Notation), a text-based open standard designed for human-readable data interchange, was used for serializing and transmitting structure data over the network by transmitting the data between a server and web application. In order to download a published study in pdf, the client URL (cUrl) was used, and Myfaces-core, Java server faces (JSF) was used when a user needs to display data on the Web. The Self-Monitoring, Analysis, and Reporting Technology (SMART) URL analysis system for

monitoring computer hard disks to detect and report on various indicators of reliability helped improve and increase system performance and provide a better user interface.

## 7.4 Implementation

### 7.4.1 Administrator Component

The administrator component (Figure 7-5) is maintained by the system administrator. For example, if a researcher is managing the crawling process, adding a new system and populating the database with studies that are specific to them, that researcher becomes an administrator. If the web crawler retrieves a study that is not relevant, the system administrator should delete it from the system database. To log into the Administrator User Interface, the user is required to obtain authentication from the system administrator. As a system administrator, the user can perform all the tasks (select study or system, view transaction, search for studies or system, modify system operations, delete or edit transactions, add system and study details).

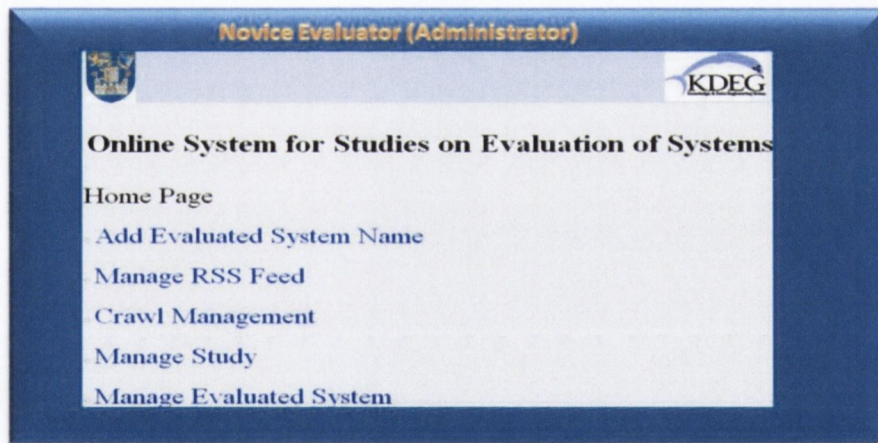


Figure 7- 5: OSSES administrator component

#### 7.4.1.1 Add System

This function allows the system administrator to quickly add, edit and view Evaluated System details (e.g. system name, functions, application area, evaluation method, criteria, system purpose).

### 7.4.1.2 RSS Feed Management

Once a system name has been added, the RSS feed management function allows the administrator to add the feed name and the URL link. It also provides a dropdown list of systems that were added in function 1 (Figure 7-6). Once the feed has been added the user is able to see the feed title and the URL link (Figure 7-7).

RSS Feed Management

FEED URL ACTION

Quick Add

RSS Feed Name: Communications of the ACM Information Systems

RSS Feed URL: http://cacm.acm.org/browse-by-subject/information-systems.rss

Evaluated System: GAS - Group Adaptive System

ADD

ADMIN HOME

Figure 7- 6: RSS feed management user interface

FEED	URL	ACTION
Communications of the ACM Information Systems	<a href="http://cacm.acm.org/browse-by-subject/information-systems.rss">http://cacm.acm.org/browse-by-subject/information-systems.rss</a>	REMOVE

Figure 7- 7: Crawled feed and the URL link

### 7.4.1.3 Crawl Management

The crawl management function performs two tasks; first, the system administrator has to crawl the studies, and then to crawl documents. This task can only be performed when the task (Manage RSS Feed) has been completed. The crawl management components are depicted in Figure 7-8.

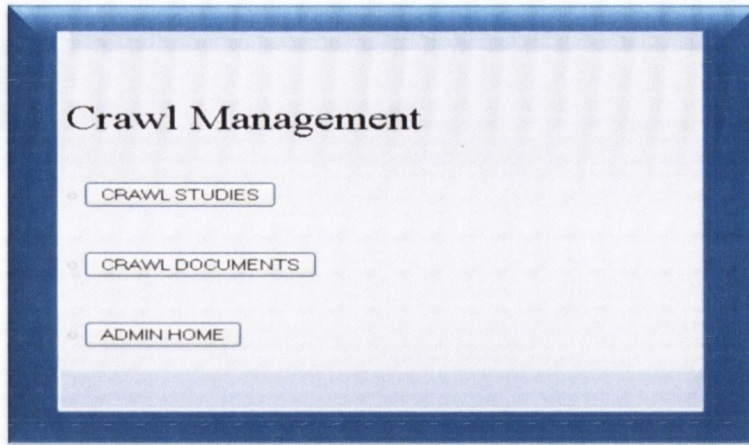


Figure 7- 8: Crawl management

#### 7.4.1.4 Study Management

For each published study, the system automatically retrieves the published study from the Web. The crawled studies are then manually sliced to create an educational evaluation dataset. Each study has a title, author, published date, link content, citation and reference.

This function allows the system administrator to view all the crawled studies from function 3 (e.g. title, authors) and perform actions such as editing the study details (e.g. title, authors, reference, citations), and it also provides a Quick Link to the dropdown list of all the Evaluated Systems (Figure 7-9).

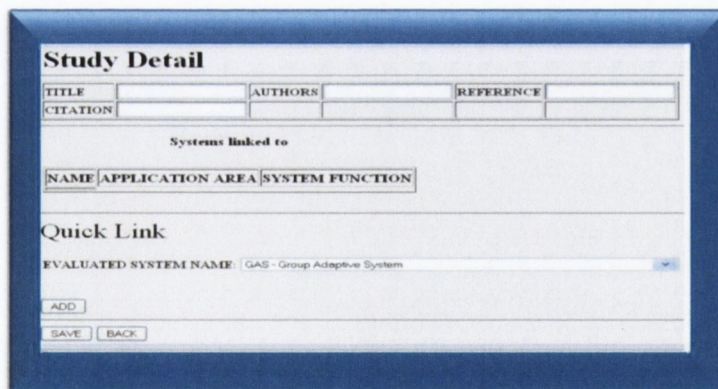


Figure 7- 9: Published study details



### 7.4.1.5 Evaluated System Management

This function allows the administrator to enter details of the system (name, function, application area, evaluation method, criteria and purpose), perform actions such as editing the system details, and also view which study is linked to that particular system. Finally, it provides details of that system (see Figure 7-10).

SYSTEM NAME	GAS - Group Adaptive	SYSTEM FUNCTION		APPLICATION AREA	
EVALUATION METHOD		CRITERIA USED		DATA TYPE ANALYSIS	
PURPOSE	Is a collaborative enviro				

Studies On: GAS - Group Adaptive System

TITLE	AUTHORS	ACTION

SAVE BACK

Figure 7- 10: Evaluated system detail

### 7.4.2 Personalized Search Component

The personalized search component is divided into five sub-components. Figure 7-11 presents a screen shot of this component.

Novice Evaluator (Administrator)

Online System for Studies on Evaluation of Systems

- Home Page
- Add Evaluated System Name
- Manage RSS Feed
- Crawl Management
- Manage Study
- Manage Evaluated System

Figure 7- 11: Personalized search component

### 7.4.2.1 Personalized Search Components

*List of evaluation studies:* The user can view the list of all studies, search a specific study, and perform actions (e.g. find details of the study such as title, author, reference, citation, systems linked to that study, system details such as system name, application area, functions).

*Search Study.* Users can search for studies and view study details (e.g. title, authors, references and citations).

*List Evaluated System.* This function provides a list of all the evaluated systems described in the published studies. It also allows users to perform actions such as: view details of the evaluated system and the studies (title and authors) describing that particular system.

### 7.4.2.2 Search Evaluated System

Each evaluated system mentioned in the retrieved studies is described in terms of: system name, the functions it fulfils, the purpose, application area, evaluation methods, criteria and metrics used.

The user can search for existing systems by using the following search terms: system name, function, application area, evaluation methods, criteria used, data type analysis, evaluated system purpose. For example, if a user is searching for a system named 'ISIS-TUTOR' (see Figure 7-12) and it exists in the database, it is displayed with all other details relating to that system (see Figure 7-12).

Search Evaluated System	
SYSTEM NAME	ISIS-TUTOR
SYSTEM FUNCTION	
APPLICATION AREA	
EVALUATION METHOD	
CRITERIA USED	
DATA TYPE ANALYSIS	
EVALUATED SYSTEM PURPOSE	
<input type="button" value="SEARCH"/>	
<input type="button" value="HOME"/>	

Figure 7- 12: Search evaluated system

## 7.5 Prototype Testing

### 7.5.1 Evaluation Process

The validation of the OSSES high fidelity prototype was subdivided into two distinct tasks: functional verification and efficiency evaluation. Functional verification was used in order to verify all the functional requirements, and the efficiency evaluation techniques to ensure user satisfaction. Different software testing elements were used:

- **Methods and techniques**, including: information retrieval techniques, interviews, expert reviews and log file production – during downloading the crawler produces a log file containing information on the pdf documents of the downloaded studies
- **Process, empirical knowledge, tools** – using the Self-Monitoring, Analysis, and Reporting Technology (SMART), which is a monitoring system for computer hard disks to detect and report on various indicators of reliability, in the hope of anticipating failures during the process of crawling.

### 7.5.3 Results and Findings

The validation of the OSSES high fidelity prototype was subdivided into two distinct tasks; functional verification and efficiency evaluation. Functional verification was used in order to verify all the functional requirements are met and efficiency evaluation techniques in order to ensure user satisfaction. Different software testing elements were used: **methods and techniques** which included; Information Retrieval techniques, interviews, expert reviews and log file production during downloading the crawler produces a log file containing some information on the PDF documents of the downloaded studies, **process, empirical knowledge, tools** the (Self-Monitoring, Analysis, and Reporting Technology (S.M.A.R.T) which is a monitoring system for computer hard disks to detect and report on various indicators of reliability, in the hope of anticipating failures during the process of crawling.

Two different types of evaluations were conducted, **formative** and **summative** evaluations. Formative evaluation was conducted during the implementation process. A range of evaluation methods were used: i) interviews with domain experts, ii) tasks and results, iii) knowledge and data engineering research group. To evaluate the crawler,

features such as completeness, robustness and download limiting and the graphical User Interface were considered. The OSSES system was evaluated internally by: i) Our research supervisors, ii) Presentation of the deployed system to a group of researchers in Knowledge and Data Engineering Group (KDEG) during our internal summer/chi workshop in college. The KDEG research group is pioneering research into the fundamental challenges and application of knowledge driven systems. The group combines innovative technology research in knowledge discovery, representation, reasoning, data management and intelligent systems engineering. The following are the questions asked by evaluators and our response:

1. *Why should we use your system while we can use Google search engine? It will save the end users time and also encourage research in the area of evaluations of systems.*
2. *Why use of different technologies and software's? Our response was these technologies and software's were used in order to increase the system performance and also for a better user interface.*
3. *What are your future plans for this system? Our future plan is to add more functions that are specific to user evaluations of adaptive systems.*
4. *How did you test the system? System was tested by functional verification and efficiency evaluations.*
5. *How do you deal with retrieved studies that are not relevant? The system administrator deletes irrelevant studies, as demonstrated by the activity diagram (see figure 4).*
6. *Can new functions be added that are relevant to our research? Yes. One of the research students wanted to start using the system immediately.*

In addition summative evaluation was conducted to provide information on the system's ability to perform better. In order to determine how well the system performed several **evaluation criteria** were used: Evaluation of input data (e.g., objectivity of data assessment, retest-reliability), Evaluation of Adaptation decision (e.g., retrieval accuracy, precision and recall, amount of help required, computational time; number of navigation steps, task success, user satisfaction), usability satisfaction, effectiveness, reliability, functionality, performance, time, robustness, downloading limiting and completeness.

### **7.6.1 System Usage**

Currently, the system contains studies and evaluated systems on user evaluation of adaptive systems, particularly adaptive systems which combine adaptive hypermedia and information retrieval techniques. For each of these studies, the relevant information is accessed, processed and recorded in the database; this information include system name, function, application area, evaluation method, criteria used, purpose of the system, and data type analysis. The evaluation of adaptive systems is not easy, and several researchers have pointed out potential pitfalls when evaluating these systems. Examples of pitfalls mentioned in (Tintarev and Masthoff, 2009b) and (Weibelzahl, 2005) include: (i) difficulty in attributing cause, (ii) insignificant results due to variance between participants, (iii) difficulty in defining the effectiveness of adaptation, (iv) allocation of insufficient resources, (v) too much emphasis on summative rather than formative evaluation and, most importantly, (vi) measures for adaptivity success have not been investigated systematically up to now.

A few researchers have implemented interactive online databases with similar functions to the OSSES, but some of these databases are out of date (Weibelzahl and Weber, 2001) making it difficult to obtain clear and up-to-date metadata for the evaluation of adaptive systems. Although, there are systems with similar databases, to our knowledge, none of the existing incorporates the focused crawling functionality. This system is a valuable tool for PhD students since it will help to reduce the cost and time required for conducting literature reviews.

## **7.6 Conclusion**

This tool support novice evaluators who are conducting literature reviews; by encouraging new researchers from different diversities to research the evaluation of systems which fulfill certain methodological requirements. It will also serve as a reference for researchers in the different fields of evaluations of any kind of system; for example research on user evaluations of adaptive systems especially those that combine adaptive hypermedia and information retrieval techniques. The online database will help to identify gaps and pitfalls in the planning process of evaluations as well as in the analysis of collected data. 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. A user who wants to use a Web crawler has two choices: building it from scratch or downloading one from the internet. The second option has some drawbacks such as the user deciding which one to choose? Which is the best for the task at hand? Which is most complete? Which is the most robust? A fully functional Web crawler which is capable of automatically retrieving recent published studies is provided. The candidate is convinced that the quality of evaluations will benefit and that, indirectly, the user will be served in the process.

## 8 Conclusion

### 8.1 Research Question & Objectives Revisited

This chapter revisits the research question and objectives of the work as stated in Chapter 1. It further discusses the findings of the user trials conducted in the course of this research. It concludes the thesis with a discussion of the overall achievements and contributions, as well as suggestions for future research direction. Specifically, section 8.2 reiterates the research question and analyses how well the research objections have been achieved. Section 8.3 discusses the overall contributions of this research and presents the research publications that have resulted from this work. Finally, section 8.4 outlines a number of future directions for evaluations of adaptive E-Learning systems and hybrid recommender TEL research.

As stated in Chapter 1 (section 1.2), this thesis investigate current evaluation techniques used by evaluators of adaptive E-Learning systems and the tradeoffs between these techniques to support user-centered evaluations of such systems. More specifically, it asks the question, *“What are the techniques used and tradeoffs between the techniques which support user-centered evaluations of adaptive systems? Can a hybrid recommender system propose appropriate evaluation methods, criteria and metrics for individual adaptive systems and to what extent are these recommendations comparable to those of human expert recommendations”*. In order to tackle the research question, it was divided it into two sub questions:

**Sub RQ1:** What are the techniques used and tradeoffs between the techniques which support user-centered evaluations of adaptive systems?

**Sub RQ2:** Can a hybrid recommender system propose appropriate evaluation methods, criteria and metrics for individual adaptive systems and to what extent are these recommendations comparable to those of human expert recommendations.

To tackle this question a hybrid of (case-study and evidence-based) approach was taken to investigate evaluation techniques used and tradeoffs between the techniques which support

user-centered evaluations of adaptive systems. This approach involved a novel combination of techniques and technologies from the areas of: i) adaptive technology enhanced learning, ii) adaptive information retrieval and iii) hybrid recommender technologies in the educational domain. The thesis resulted in the specification, design and development of an evaluation framework for supporting novice and expert evaluators of adaptive systems (EFEx), which consists of three major components component:

- (i) An automated hybrid (case-based and knowledge-based) system for recommending evaluation techniques. This approach overcomes the limitations of case-based and knowledge-based recommendation techniques as defined in Chapter 2 Section 2.3.3.5. The multi-attribute relationships that need to be traversed by humans to work out the most appropriate evaluation techniques are not easily navigated using typical database techniques. Recommendation technology was used, therefore, to enhance the appropriateness of suggested evaluation techniques.
- (ii) A personalised search component that allows users to find evaluation studies of adaptive systems and
- (iii) A taxonomy of technical terms for supporting the evaluation of adaptive systems. This taxonomy will assist novice evaluators using the recommender system.

The architecture and implementation of EFEx framework is discussed in chapters 3 and 4. Elicitation of knowledge-base for EFEx is discussed in chapter 5. Furthermore chapter 6 presents evaluation results and findings of EFEx. In addition chapter 7 discusses OSSES focused online crawling system for evaluation studies of adaptive systems.

A series of EFEx-based high-fidelity prototypes were developed and evaluated using several software testing methodologies and techniques. These showed that the hybrid recommendation service developed can recommend a combination of evaluation techniques more accurately and appropriately compared to a human expert. Furthermore, the approach enabled us to identify existing evaluation techniques used and tradeoffs between these techniques to support user-centered evaluations of adaptive systems.

In addition, to addressing the overall research question:



*“What are the techniques used and tradeoffs between the techniques which support user-centered evaluations of adaptive systems? Can a hybrid recommender system propose appropriate evaluation methods, criteria and metrics for individual adaptive systems and to what extent are these recommendations comparable to those of human expert recommendations”.* This thesis has also met the individual research objectives identified in chapter 1 (section 1.3). Each of these objectives is discussed below.

**Research Objective1** *Investigate what are the capabilities and tradeoffs that user-centered evaluation (UCE) techniques can discover or estimate; through literature and survey.*

The first objective was achieved by conducting an extensive review that investigated evaluation techniques used and tradeoffs between those techniques to support user-centered evaluations in the field of adaptive systems. Furthermore, a real life user study survey on evaluations of adaptive systems developed from 2000 to 2011 was conducted and the results analysed (chapter 5, section 5.2.3). Based on these results an educational evaluation dataset was created (chapter 5, section 5.3). This dataset was used to populate EFEx database.

One of the evaluation approaches which was applied when tackling this objective was to collect evidence through a real-life study. The study enabled the identification of respective tradeoffs that user-centered evaluation techniques could discover or estimate which are discussed in Chapter 7, section 7.2. The study enabled us to identify respective tradeoffs between different evaluation techniques (Chapter 2 Section 2.5).

The achievement of objective 1 is presented in the following publications:

- Mulwa, C., Lawless, Sharp, M. and Wade, V. (2011). The Evaluation of Adaptive and User Adaptive Systems: A Review. *International Journal of Knowledge and Web Intelligence (IJKWI)*, pp 138-156.
- Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I., and Wade, V. (2010). Adaptive educational hypermedia systems in technology enhanced learning: a literature review. *Proceedings of the 2010 ACM Conference on Information Technology Education*, pp 73-84.

- Mulwa, C., McDonald, H., O'Keeffe, I., Lewis, D., He, Y., and Wade, V. (2012). The Maturity Model: A Novel Way of Evaluating Centre for Next Generation Localisation Demonstrator Systems Based on Industrial Impact, Scientific Collaboration and Interoperability. *Proceedings of World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE), pp 1340-1349.
- Mulwa, C., Li, W., Lawless, S., and Jones, G. (2010). A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction. *In Proceedings of the Workshop on Simulation of Interaction. Automated Evaluation of Interactive IR , SIGIR 2010*.
- Lawless, S., O'Connor, A. and Mulwa, C. (2010). A Proposal for the Evaluation of Adaptive Personalised Information Retrieval", *Proceedings of the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation held in conjunction with ECIR2010 European Conference on Information Retrieval*.

**Objective2**     *Design, develop and evaluate a focused crawling system for evaluation studies of adaptive systems*

The **second objective** was realised through the development of a web-based focused crawling system that uses several crawling techniques to automatically crawl and present to the user relevant evaluation studies of adaptive systems. This tool assists novice users who are investigating the literature in the area of evaluation of adaptive systems. The author has used the system to crawl evaluation studies published since 2000. To date over 450 published studies have been crawled and manually analyzed. The resulting educational evaluation dataset (section 7.3) was used to populate the EFEx evaluation framework developed under objective 3.

The achievement of objective 2 is presented in the following publication:

- Mulwa, Mulwa, C., Lawless, S., Sharp, M. and Wade, V. (2010). OSSES: An Online System for Studies on Evaluation of Systems, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Association for the Advancement of Computing in Education (AACE), pp 3210-3219.

The third objective, which represents a key contribution in this research, realised through the development of an EEx evaluation framework

**Objective3**      *Design and develop an evaluation framework for supporting novice and expert evaluators of adaptive systems (EEx); which consists of three components major component:*

- (i) An automated hybrid (case-based and knowledge-based) system for recommending evaluation techniques,*
- (ii) A personalised search component that allows users to find evaluation studies of adaptive systems and*
- (iii) A taxonomy of technical terms for supporting the evaluation of adaptive systems )*

**Research objective 3 (i)** was realised through the implementation of an automated, novel, hybrid (case-based and knowledge-based) system for recommending evaluation techniques to novice and expert evaluators of adaptive systems. Currently the system is focused on recommending a combination of evaluation techniques for adaptive E-Learning systems (AELS). To the candidates' knowledge, there are no other similar hybrid recommender systems that exist. The evaluation of AELS is a difficult task due to the complexity of such systems, as shown by many studies (Chapter 2). 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. Many metrics can be used to measure performance, for example: knowledge gain (AELS), amount of requested materials, duration of interaction, number of navigation steps, task success, usability (e.g. effectiveness, efficiency and user satisfaction). The evaluation of adaptive systems is not easy, and several researchers have pointed out potential pitfalls. Further, evaluators have emphasized the difficulty in ensuring that the correct evaluation techniques are used in order to produce accurate results and solve the

usability issues associated with such systems. This thesis would lead us to believe this new system is a valuable tool and assists novice evaluators of adaptive systems.

**Research objective 3 (ii)** was realised through the implementation of a personalized search system component that allows users to find evaluation studies of adaptive systems (specifically, adaptive E-Learning systems) and evaluations of the internal models of such systems (discussed in section 4.4.2). This system assists novice evaluators interested in finding out how adaptive systems developed since 2000 were evaluated. Furthermore, researchers are presented with a centralized database and an automated interactive user interface, divided into three distinct user interfaces that allow them to find evaluation studies of (i) the internal models of adaptive systems, (ii) adaptive systems developed and evaluated between 2000 and 2013, and (iii) general evaluation studies of such systems.

Research **objective 3 (iii)** was realised through the production of a taxonomy of technical terms for supporting the evaluation of adaptive systems. This taxonomy consists of a standardized categorization of learning objectives (terms) in an educational context, aimed at increasing understanding of terminologies used by evaluators of adaptive systems. Furthermore, it provides basic understanding about the components of evaluations of AEL systems in general.

***Objective 4 Evaluate the three components of the evaluation framework designed in Objective3***

Finally, **Objective 4** was realised by conducting a series of evidence-based evaluations that consisted of: i) structured interviews, ii) task-based, real user studies, and iii) online structured questionnaires. These were chosen because they are effective methods for measuring accuracy and appropriateness of recommendations, user satisfaction, learnability and usability. The analyzed results presented are presented in Chapters 4, 6 and 7, respectively; also show that both communities acknowledge the importance of such systems, especially the recommender system.

The initial evaluation was part of the design but also addressed user satisfaction. The results and findings of these evaluations are discussed in Chapter 4 Section 4.3.3. Most participants (novice and expert evaluators) suggested that the recommendation service and the personalized search component were of significant use to them. These results were

used to improve the initial high-fidelity prototype developed. It was important to collect evidence from real users on which components of EEx they considered useful. The results of this evaluation enabled us to tailor the systems specifically for our target user group and to provide a tool that would be useful to them.

The second evaluation was part of implementation which focused on system-specific function testing techniques and validating the recommendation appropriateness of the algorithm but also addressed the perceived usefulness and usability from the user's perspective (user satisfaction) of the prototype. After interacting with the EEx, 85.2% of participants (experts and novice evaluators) agreed they found the developed components of the EEx framework useful. The results of these evaluations and findings are discussed in Chapter 4 Section 4.4.4. The participants' opinion is correct because, these evaluators were the correct targeted group who were developers and also evaluators adaptive systems developed from 2000 to 2014.

Finally, a third round of evaluations of all the components of EEx framework (chapter 6, section 6.2) – which consisted of a combination of evaluation techniques: (i) user-centred evaluations, (ii) task-based techniques, (iii) structured interviews, (iv) post questionnaires and recommended evaluation techniques appropriateness. These evaluations addressed usability (user satisfaction), retrieval identification, and appropriateness of recommended evaluation techniques, learnability and perceived usefulness of all EEx components. The results and findings are discussed in Chapter 6 Sections 6.2 to Section 6.4. The results confirmed that the developed framework can successfully support novice evaluators of adaptive systems, especially in getting appropriate recommendations on evaluation approaches and a combination of evaluation techniques (methods, metrics and criteria). Furthermore, provision of such a tool can improve end-user experience and learnability during evaluations of such systems. Thus answering (Sub RQ2)

In total, the various prototype implementations have been evaluated by almost 248 participants, who were both developers and evaluators of adaptive systems. Of these, 50 were domain expert evaluators of such systems from the E-Learning, recommender and adaptive hypermedia communities. The evaluation studies and experiments have each confirmed that the developed framework resulting from this research can accurately and appropriately recommend evaluation approaches and a combination of evaluation techniques to novice and expert evaluators of adaptive systems. Furthermore, the

centralized database composed of an educational evaluation dataset is useful to both recommender and AEL communities.

The achievement of objectives 3 and 4 is presented in the following publications:

- Mulwa, C., and Wade, V. (2013). A Web-Based Evaluation Framework for Supporting Novice and Expert Evaluators of Adaptive E-Learning Systems. *Proceedings of International Conference on E-Technologies and Business on the Web (EBW2013)*, Society of Digital Information and Wireless Communication, pp. 62-67.
- Mulwa, C. Lawless, S., O’Keeffe, I., Sharp, M., and Wade, V. (2012). A Recommender Framework for the Evaluation of End User Experience in Adaptive Technology Enhanced Learning”. *International Journal of Technology Enhanced Learning (IJTL)*, pp. 67-84.
- Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2012). The Evaluation of Adaptive Technology Enhanced Learning Systems. *Proceedings of World Conference on E-Learning in Corporate, Government, and Healthcare and Higher Education (ELEARN)*, Association for the Advancement of Computing in Education (AACE), pp 744-753.
- Mulwa, C. Lawless, S., Sharp, M., and Wade, V. (2011). A Web based Framework for the Evaluation of End User Experience in Adaptive and Personalised E-Learning Systems. *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE Computer Society, vol. 3, pp 351-356
- Mulwa, C., Lawless, S., Ghorab, M.R., O'Donnell E, Sharp, M., and Wade, V. (2011). A framework for the evaluation of adaptive IR systems through implicit recommendation. In: Andrews, S., Polovina, S., Hill, R., Akhgar, B. (Eds.) *Conceptual Structures for Discovering Knowledge*. Springer Berlin Heidelberg. pp 366-374
- Mulwa, C., Lawless, S., Sharp, M., Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems a (Demonstration Paper). *Proceedings of User Modeling, Adaptation and Personalization Conference, (UMAP 2011)*.
- Mulwa, C., Lawless, S., Sharp, M., Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems, a Demonstration Paper.

*Proceedings of User Modeling, Adaptation and Personalization Conference, (UMAP 2011).*

- Mulwa, C., Longo, L., Lawless, S., Sharp, M., and Wade, V. (2011). An online framework for supporting the evaluation of personalised information retrieval systems. *Proceedings of the 6th international conference on Ubiquitous and Collaborative Computing (UCC 2011)*. British Computer Society. pp 75-85
- Mulwa, C., Li, W., Lawless, S., and Jones, G. (2010). A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction. *In Proceedings of the Workshop on Simulation of Interaction: Automated Evaluation of Interactive IR (SIGIR 2010).*

The evaluations of the novices seem to prove that both the evaluations made by the hybrid recommender system seem credible, well argued and well backed up and therefore majority of percentage of them said the recommended evaluation approaches to be used and techniques seemed to be correct. Unfortunately the candidate cannot guarantee that the novices are the right people to judge whether the recommendations produced are correct.

The caveats of this thesis are that it is difficult to make strong claim that the hybrid recommender system is as good as what would be recommended by the human experts. One of the reasons is because it is difficult to get enough experts and also we found it difficult to make complete comparable tests.

Through a series of evidence-based case studies and structured interviews, this thesis has provided a real picture of what people are doing in evaluations of adaptive systems developed between 2000; more specifically adaptive E-Learning systems. The results of these studies are analyzed in Chapters 4, 5, 6 and 7 respectively. The overall findings after evaluating EFEx framework and the focused crawling system has shown that the four research objectives has been met successfully, as users were able to, from: i) an accuracy perspective; get appropriate recommendations of evaluation techniques, approaches to use and explanations on how the techniques were derived, ii) a usability perspective (user satisfaction), after interacting with the recommender, personalised search system and the taxonomy users found the systems useful and iii) from learnability perspective, novice users were able to learn after interacting with the framework how evaluations have been conducted since 2000 (addressed in Chapter 6 Section 6.4 and Chapter 7 Section 7.2 and 7.3).

The candidate can argue, the results and findings of the both novice and expert evaluators after completing experiment 1, 2 and 3 (research objective 4) are correct because the results seem to prove that both the evaluations produced by EFEx framework seem credible, well argued and well backed up (chapter 4, section 4.4.4 and chapter 6).

## **8.2 Contributions to the research field**

As specified in section 1.5, this research has resulted in one major contribution and two minor contributions. In addition to the research work presented, 17 publications have resulted from this work s (see Appendix A).

The major contribution of this thesis is a novel hybrid (case-based and knowledge-based) recommender system for recommending evaluation techniques and approaches. In particular, this innovative recommendation service can recommend appropriate and accurate evaluation approaches to be used during evaluations of adaptive systems. In addition evaluators can get a combination of evaluation techniques (methods, criteria and metrics). The recommended approaches and techniques are ranked and the top five most accurate techniques presented to the user. The recommendation service also acknowledges the hard work of expert (three years and above) evaluators, and provides them with an extra functionality which allows them to ignore the recommended evaluation approaches and allows them to choose their own from a list of approaches identified in the literature. To the candidates knowledge there is no other evidence-based study that provides a real picture of what people are doing in evaluations of adaptive systems. Furthermore, there are no other automated hybrid recommender systems (educational domain) in the field of adaptive systems. Thus a significant contribution is made to the body of knowledge.

In addition, the thesis has presented two architectures; an Evaluation Framework for Supporting Evaluators of Adaptive Systems (EFEx) and a focused crawling System for Evaluation Studies (OSSES). The EFEx architecture implements this recommendation service by combining case-based and knowledge-based recommendation capabilities. In addition it implements a personalised search system for evaluations studies of adaptive systems developed from 2000 and a taxonomy of technical terms to support the search. The two architectures have been used as the basis of two prototypes (presented in Chapters



4 and 5), which have been acknowledged by both recommender and AEL communities as being useful.

This research has resulted in a number of high-quality scientific publications, which are briefly discussed next.

The analysis of evaluations of adaptive E-Learning systems developed from 2000 to 2013 (presented in Chapter 2) has been published in the *International Journal of Knowledge and Web Intelligence (IJKWI)*. In addition it has been resulted in publication of 2 full conference papers which were presented at the *ACM Special Interest Group for Information Technology Education Conference (SIGITE)* and the *World Conference on Educational Media and Technology (EDMEDIA)*. These three papers reviewed current evaluation techniques to support UCE evaluations of adaptive systems, more specifically AEHS systems. A review of current evaluation challenges, pitfalls and difficulties encountered by evaluators of such systems was also conducted.

- Mulwa, C., Lawless, Sharp, M. and Wade, V. (2011). The Evaluation of Adaptive and User Adaptive Systems: A Review. *International Journal of Knowledge and Web Intelligence(IJKWI)*, pp 138-156
- Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I., and Wade, V. (2010). Adaptive educational hypermedia systems in technology enhanced learning: a literature review. *Proceedings of the 2010 ACM Conference on Information Technology Education*, pp 73-84.
- Mulwa, C., McDonald, H., O'Keeffe, I., Lewis, D., He, Y., and Wade, V. (2012). The Maturity Model: A Novel Way of Evaluating Centre for Next Generation Localisation Demonstrator Systems Based on Industrial Impact, Scientific Collaboration and Interoperability. *Proceedings of World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE), pp 1340-1349.
- Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2012). The Evaluation of Adaptive Technology Enhanced Learning Systems. *Proceedings of World Conference on E-Learning in Corporate, Government, and Healthcare and Higher Education (ELEARN)*, Association for the Advancement of Computing in Education (AACE), pp 744-753.

Having reviewed current evaluations of adaptive E-Learning systems, the candidate further published two workshop papers which were presented at the *33rd Annual ACM SIGIR Conference (2010)* and the *European Conference on Information Retrieval (2010)*. Both papers proposed a new approach to the evaluation of adaptive information retrieval systems (AIR):

- Mulwa, C., Li, W., Lawless, S., and Jones, G. (2010). A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction. In *Proceedings of the Workshop on Simulation of Interaction: Automated Evaluation of Interactive IR (SIGIR 2010)*, ACM.
- Lawless, S., O'Connor, A. and Mulwa, C. (2010). A Proposal for the Evaluation of Adaptive Personalised Information Retrieval", *Proceedings of the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation held in conjunction with ECIR2010 European Conference on Information Retrieval*.

Furthermore the EFEx architecture and prototypes (presented in Chapter 4) and their application potential in supporting novice and expert evaluation of adaptive systems were published in the *International Journal of Technology Enhanced Learning (2012)*. This paper specifically focused on the ability of the automated hybrid recommender system in recommending appropriate and accurate evaluation techniques.

- Mulwa, C. Lawless, S., O'Keeffe, I., Sharp, M., and Wade, V. (2012). A Recommender Framework for the Evaluation of End User Experience in Adaptive Technology Enhanced Learning". *International Journal of Technology Enhanced Learning (IJTL)*, pp. 67-84.

Reports of the implementation and evaluation of EFEx prototypes were published in several conferences, lecture notes in artificial intelligence (LNAI) and workshops. The results and findings of task-based evaluations of the framework were published at the *International Conference on E-Technologies and Business on the Web*.

- Mulwa, C., and Wade, V. (2013). A Web-Based Evaluation Framework for Supporting Novice and Expert Evaluators of Adaptive E-Learning Systems. *Proceedings of International Conference on E-Technologies and Business on the Web (EBW2013)*, Society of Digital Information and Wireless Communication, pp. 62-67.
- Mulwa, C. Lawless, S., Sharp, M., and Wade, V. (2011). A Web based Framework for the Evaluation of End User Experience in Adaptive and Personalised E-

Learning Systems. *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE Computer Society, vol. 3, pp 351-356

- Mulwa, C., Lawless, S., Ghorab, M.R., O'Donnell E, Sharp, M., Wade, V. (2011). A framework for the evaluation of adaptive IR systems through implicit recommendation. In: Andrews, S., Polovina, S., Hill, R., Akhgar, B. (Eds.) *Conceptual Structures for Discovering Knowledge*. Springer Berlin Heidelberg. pp 366-374.
- Mulwa, C., Longo, L., Lawless, S., Sharp, M., and Wade, V. (2011). An online framework for supporting the evaluation of personalised information retrieval systems. *Proceedings of the 6th international conference on Ubiquitous and Collaborative Computing, 2011*. British Computer Society. pp 75-85.

In addition the candidate published a demonstrator and poster paper at the *user modeling, adaptation and personalization (UMAP) conference*. The EEx prototype was demonstrated to the conference attendants. Most of the researchers who visited our demonstration section acknowledged, after interacting with the automated version of the EEx framework, that they found it useful. The results of the structured questionnaire filled out by participants are presented in Chapter 7.

- Mulwa, C., Lawless, S., Sharp, M., Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems a (Demonstration Paper). *Proceedings of User Modeling, Adaptation and Personalization Conference, (UMAP 2011)*.

The design, implementation and evaluation of the focused crawling system currently been used by the author to crawl evaluation studies of AEL systems (presented in Chapter 5) have been published in a full conference paper at the World Conference on Educational Multimedia, Hypermedia and Telecommunications (2010). This paper described the functionalities of the developed OSSE system and potential benefits of such a system to end users.

- Mulwa, C., Lawless, S., Sharp, M. and Wade, V. (2010) "OSSES: An Online System for Studies on Evaluation of Systems", *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Association for the Advancement of Computing in Education (AACE), pp 3210-3219.

In conclusion, the candidate believes that this research is the first evidence-based research that provides a real picture of who is doing what in order to evaluate what, in the field of evaluation of Adaptive E-Learning Systems.

### **8.3 Future Research Suggestions**

The research described in this thesis has resulted in a novel hybrid recommendation service for AEL systems, a personalized search interface, a taxonomy of technical terms to help support the search, and a focused system for crawling evaluation studies. An evaluation educational dataset was also created and currently being used to populate the database of EFEx framework (discussed in chapter 7). Building on these encouraging results, there are a number of opportunities for future research.

#### ***UCE evaluations of the internal models of AEL***

The results of the evidence-based study (section 7.2) show that very limited research has been conducted in evaluations of the inner metadata models of adaptive systems (see section 2.3.2). Due to the complexity of such systems and the usability issues encountered by users, it is important that proper evaluations of these models be conducted during and after implementation. These evaluations might provide solutions to the current usability issues.

#### ***Evaluation educational dataset for AEL systems***

An investigation of the literature on existing evaluation educational datasets showed that there are not enough evaluation educational datasets for recommender systems in the E-Learning domain. These datasets are important in order to improve and increase the performance of evaluation of educational recommender systems.

Several researchers have noted the need for datasets 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 datasets that can be adopted and reused by the scientific community. Datasets 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 datasets that cover the spectrum of formal and informal learning.

## Bibliography

- Akilli, G. K. (2007). Games and Simulations: A New Approach in Education?" in Games and Simulations in Online Learning: Research and Development Frameworks, D. Gibson, C. Aldrich, and M. Prensky, Eds.: Information Science Pub, 2007, pp. 1-20.
- Aroyo, L., Mizoguchi, R., and Tzolov, C. (2003). OntoAIMS: Ontological Approach to Courseware Authoring, pp. 2-5.
- Breitner, M. and Hoppe, G. (2005). A Glimpse at Business Models and Evaluation Approaches for E-Learning, pp. 179-193.
- Brown, E. (2007). The use of learning styles in adaptive hypermedia. Phd, University of Nottingham.
- Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. User modeling and user-adapted interaction, 6, pp. 87-129.
- Brusilovsky, P. (2001). Adaptive hypermedia. User Modeling and User-Adapted Interaction, vol. 11(1-2), pp. 87-110.
- Brusilovsky, P. (2004). Adaptive Educational Hypermedia: From Generation to Generation, pp. 19-33.
- Brusilovsky, P. (2004). Adaptive Navigation Support: From Adaptive Hypermedia to the Adaptive Web and Beyond. PsychNology Journal, vol. 2, pp. 7-23.
- Brusilovsky, P. (2004). KnowledgeTree: A distributed Architecture for Adaptive e-Learning, pp. 104-113.
- Brusilovsky, P., Chavan, G., and Farzan, R. (2004). Social Adaptive Navigation Support for Open Corpus Electronic Textbooks. Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2004, LNCS.
- Brusilovsky, P., Karagiannidis, C. and Sampson, D. (2004). Layered evaluation of adaptive learning systems. International Journal of Continuing Engineering Education and Life Long Learning, vol. 14, pp. 402-421.
- Brusilovsky P., Karagiannidis C. and D., S. (2001). The Benefits of Layered Evaluation of Adaptive Applications and Services. Proceedings of the First Workshop on Empirical Evaluation of Adaptive Systems, pp.1-8.
- Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, vol. 12, pp. 331-370.

- Chin, D. 2001. Empirical Evaluation of User Models and User-Adapted Systems. *User Modeling and User-Adapted Interaction*, vol. 11, pp. 181-194.
- Conlan, O. 2005. *The Multi-model, Metadata Driven Approach to Personalised eLearning Services*. PhD, Trinity College.
- Conlan, O., Hampson, C., Peirce, N., and Kickmeier-Rust, M. (2009). Realtime Knowledge Space Skill Assessment for Personalized Digital Educational Games. *Advanced Learning Technologies*. Ninth IEEE International Conference, pp. 538-542.
- Conlan, O., and Wade, V. 2004. Evaluation of APeLS - An Adaptive eLearning Service Based on the Multi-model, Metadata-driven Approach. *Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems Proceedings*.
- Conlan, O., Wade, V., Bruen, C. and Gargan, M. (2002). Multi-model, Metadata Driven, Approach to Adaptive Hypermedia Services for Personalized eLearning. *Adaptive Hypermedia and Adaptive Web-Based Systems*, pp. 100-111.
- Csikszentmihalyi M. (1990). *Flow: The psychology of optimal experience*. NY: New York: Harper and Row.
- Dagger, D., Wade, V., and Conlan, O. (2005). Personalisation for all: Making Adaptive Course Composition Easy. *Educational Technology and Society*, vol. 8, pp. 9-25.
- Davies, P. (1999). *The Virtual School of Biodiversity: Towards a Model for Quality Assured Distributed Learning*. *Proceedings of the Fifth Hong Kong Web Symposium*.
- De bra, P. (2002). Adaptive Educational Hypermedia on the Web. *Communications of the ACM*, pp. 45-61.
- De Bra, P., Aerts, A., Berden, B., De Lange, B., Rousseau, B., Santic, T., Smits, D. and Stash, N. (2003). AHA! The Adaptive Hypermedia Architecture. *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia*, pp. 81-84.
- De Bra, P., Aerts, A., Smits, D. and Stash, N., (2002). AHA! Version 2.0, More Adaptation Flexibility for Authors, pp. 240-246.
- De Bra, P. and Calvi, L. (1998). AHA! An Open Adaptive Hypermedia Architecture. *New Review of Hypermedia and Multimedia*, vol. 4, pp. 115-139.
- De Jong, M. and Schellens, P. (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 and Management*, vol. 44, pp. 1293-1307.
- Dolog, P., Henze, N., Nejdl, W. and Sintek, M., (2004). The Personal Reader: Personalizing and Enriching Learning Resources Using Semantic Web Technologies. *Third International Adaptive Hypermedia and Adaptive Web-based Systems Conference*, vol. 2, pp. 85-94.
- Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., Beham, G., Lindstaedt, S., Stern, H. and Friedrich, M. (2010). Issues and Considerations Regarding Sharable Data Sets for Recommender Systems in Technology Enhanced Learning. *Procedia Computer Science*, vol. 1, pp.2849-2858.
- Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D. and Leifer, L. J. (2005). Engineering Design Thinking, Teaching, and Learning. *Journal of Engineering Education*, vol. 94, pp.103-120.
- Ehlers, U. D., Goertz, L., Hildebrandt, B. and Pawlowski, J. M. (2005). Quality in e-Learning: Use and Dissemination of Quality Approaches in European e-Learning: a Study by the European Quality Observatory, Office for Official Publications of the European Communities.
- Esichaikul, V., Lamnoi, S. and Bechter, C. (2011). Student Modelling in Adaptive E-Learning Systems. *Knowledge Management and E-Learning: An International Journal (KM and EL)*, vol. 3, pp. 342-355.
- Farooq, A., Dumke, Reiner R. (2008). *Evaluation Approaches in Software Testing*. Faculty of Computer Science, University of Magdeburg.
- Frankola, K. (2001). Why Online Learners Drop Out. *Workforce-Costa Mesa.*, vol. 80, pp. 52-61.
- Fu , L., Salvendy , G., and Turley , L. (2002). Effectiveness of User Testing and Heuristic Evaluation as a Function of Performance Classification. *Behaviour and Information Technology*, vol. 21, pp. 137-143.
- Gena, C. (2005). *Methods and Techniques for the Evaluation of User-Adaptive Systems*. The Knowledge Engineer Review , United Kingdom: Cambridge University Press. vol. 20, pp. 1-37.
- Gena, C., and Weibelzahl, S. (2007). Usability Engineering for the Adaptive Web. *The Adaptive Web*, vol. 4321, pp. 720-762.



- Girardi, C., Ricca, F. and Tonella, P. (2006). Web Crawlers Compared. *International Journal of Web Information Systems*, vol. 2, pp. 85-94.
- Gupta, A. and Grover, P., (2004). Proposed Evaluation Framework for Adaptive Hypermedia Systems.
- Herder E., (2003). Utility-Based Evaluation of Adaptive Systems. *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems*, at the 9th International Conference on User Modeling, pp. 25-30.
- Hernández Del Olmo, F. and Gaudioso, E. (2008). Evaluation of Recommender Systems: A New Approach. *Expert Systems with Applications*, vol. 35, pp.790-804.
- Höök, K. (1997). Evaluating the Utility and Usability of an Adaptive Hypermedia System. *Proceedings of the 2nd International Conference on Intelligent User Interfaces*, pp. 179-186.
- Jaakkola H. and B., T. (2011). Architecture-Driven Modelling Methodologies. *Proceedings of the Conference on Information Modelling and Knowledge Bases XXII*. Anneli Heimbürger et al.(eds). IOS Press.
- Jameson, A. (2001). *Systems That Adapt to Their Users: An Integrative Perspective*. Saarbrücken: Saarland University.
- Jameson, A. (2009). Adaptive Interfaces and Agents. *Human-Computer Interaction: Design Issues, Solutions, and Applications*, pp. 105.
- Jannach, D., Zanker, M., Felfernig, A. and Friedrich, G. (2010). *Recommender Systems An Introduction*, Cambridge University Press.
- Johnson, W. L., Wang, N. and Wu, S., (2007) Experience With Serious Games for Learning Foreign Languages and Cultures. *SimTecT Conference*.
- Jones, A. (1996). The Use of Computers to Support Learning in Children with Emotional and Behavioural Difficulties. *Computers and Education*, vol. 26, pp. 81-90.
- Jovanovica, J., Gas Evicb, D., Torniaic, C., Batemand, S. and Hatalae, M. (2009). The Social Semantic Web in Intelligent Learning Environments: State of the Art and Future Challenges. *Interactive Learning Environments*, vol. 17, pp. 273-309.
- Karagiannidis C. and Sampson D., (2000). Layered Evaluation of Adaptive Applications and Services. *International Conference on Adaptive Hypermedia and Adaptive Applications and Services*.
- Karampiperis, P. and Sampson, D. (2005). Adaptive Learning Resources Sequencing in Educational Hypermedia Systems. *Educational Technology and Society*, vol. 8, pp. 128-147.

- Kickmeier-Rust, M. D., Peirce, N., Conlan, O., Schwarz, D., Verpoorten, D. and Albert, D. (2007). Immersive Digital Games: The Interfaces for Next-generation e-Learning? Universal Access in Human-Computer Interaction. Applications and Services.
- 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.
- Koedinger, K. R., Anderson, J. R., Hadley, W. H. and Mark, M. A. (1997). Intelligent Tutoring Goes to School in the Big City. *International Journal of Artificial Intelligence in Education (IJAIED)*, vol. 8, pp. 30-43.
- Koper, R. and Olivier, B. (2003). Representing the Learning Design of Units of Learning. *Educational Technology and Society*, vol. 7, pp. 97-111.
- Kravicik, M. and Specht, M. (2004). Flexible Navigation Support in the Winds Learning Environment for Architecture and Design, pp.156-165.
- 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., O'Connor, A., and Mulwa, C. (2010). A Proposal for the Evaluation of Adaptive Personalised Information Retrieval, Proceedings of the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation Held in Conjunction with European Conference on Information Retrieval, ECIR 2010.
- Lewis, J. R., and Sauro, J. (2009). The Factor Structure of the System Usability Scale. Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009, pp. 94-103.
- Liu F., Tong J., Bohn, J., Messina, J. and L., B.(2011). NIST Cloud Computing Reference Architecture. [http://www.nist.gov/manuscript-publication-search.cfm?pub\\_id=909505](http://www.nist.gov/manuscript-publication-search.cfm?pub_id=909505).
- Lorenzi12, F., Dos Santos, D. S. and Bazzan, A. L. (2005). Case-Based Recommender System Inspired by Social Insects.
- Magoulas, G., Chen, S. and Papanikolaou, K. (2003). Integrating Layered and Heuristic Evaluation for Adaptive Learning Environments. Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, 9th International Conference on User Modelling.
- Magoulas, G. D., Chen, S. Y. and Dimakopoulos, D. (2004). A Personalised Interface for Web Directories Based on Cognitive Styles. *User-Centered Interaction Paradigms for Universal Access in the Information Society*.

- 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.
- Malone, T. W. (1981). Toward a Theory of Intrinsically Motivating Instruction. *Cognitive Science*, vol. 5, pp. 333-369.
- Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., and Koper, R. (2011). Recommender Systems in Technology Enhanced Learning. *Recommender Systems Handbook*.
- Mansar, S. L., Marir, F., and Reijers, H. A. (2003). Case-based Reasoning as a Technique for Knowledge Management in Business Process Redesign. *Electronic Journal on Knowledge Management*, vol. 1, pp.113-124.
- 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, pp. 16–17.
- Masthoff, J. (2006). The User as Wizard: A Method for early Involvement in the Design and Evaluation of Adaptive Systems.
- Merrill, M. D. (2002). First Principles of Instruction. *Educational Technology Research and Development*, vol. 50, pp. 43-59.
- Missier Del, F. and Ricci, F. (2003). Understanding Recommender Systems: Experimental Evaluation Challenges, pp. 31-40.
- Mitchell, A., and Savill-Smith, C. (2004). The Use of Computer and Video Games for Learning: A Review of the Literature. Learning and Skills Development Agency, London.
- Mitrovic, A., Martin, B., and Suraweera, P. (2007). Intelligent Tutors for All: Constraint-based Modeling Methodology, Systems and Authoring. *IEEE Intelligent Systems*, vol. 22(4), pp. 38-45.
- Mitrovic, A., Mcguigan, N., Martin, B., Suraweera, P., Milik, N., and Holland, J. (2008). Authoring Constraint-based Tutors in ASPIRE: A Case Study of a Capital Investment Tutor. Vienna, Austria: World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 4607-4616.
- Mitrovic, A. and Ohlsson, S. (1999). Evaluation of a Constraint-based Tutor for a Database Language. *Artificial Intelligence in Education* vol. 10 (3-4), pp. 238-256.

- Mitrovic, A., Suraweera, P., Martin, B., and Weerasinghe, A. (2004). DB-suite: Experiences with Three Intelligent, Web-based Database Tutors. *Journal of Interactive Learning Research*, vol. 15, pp. 409-432.
- Mu, X., Ryu, H., and Lu, K. (2011). Supporting Effective Health and Biomedical Information Retrieval and Navigation: A Novel Facet View Interface Evaluation. *Journal of Biomedical Informatics*, vol. 44, pp. 576-586.
- Mulwa, C., McDonald, H., O'Keeffe, I., Lewis, D., He, Y., and Wade, V. (2012). The Maturity Model: A Novel Way of Evaluating Centre for Next Generation Localisation Demonstrator Systems Based on Industrial Impact, Scientific Collaboration and Interoperability. *Proceedings of World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education, AACE 2012, pp. 1340-1349.
- Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I., and Wade, V. (2010). Adaptive Educational Hypermedia Systems in Technology Enhanced Learning: A Literature Review. *Proceedings of the 2010 ACM Conference on Information Technology Education*, pp. 73-84.
- Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2010). OSSES: An Online System for Studies on Evaluation of Systems, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Association for the Advancement of Computing in Education (AACE), pp. 3210-3219.
- Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2011). The Evaluation of Adaptive and User Adaptive Systems: A Review. *International Journal of Knowledge and Web Intelligence (IJKWI)*, vol. 20, pp. 138-156.
- Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems a (Demonstration Paper). *Proceedings of User Modeling, Adaptation and Personalization Conference, UMAP 2011*.
- Mulwa, C., Lawless, S., O'Keeffe, I., Sharp, M., and Wade, V. (2012). A Recommender Framework for the Evaluation of End User Experience in Adaptive Technology Enhanced Learning. *International Journal of Technology Enhanced Learning*, pp. 67-84.
- Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2012). The Evaluation of Adaptive Technology Enhanced Learning Systems. *Proceedings of World Conference on E-Learning in Corporate, Government, and Healthcare and Higher Education, ELEARN 2012*, Association for the Advancement of Computing in Education AACE, pp. 744-753.

- Mulwa, C. Lawless, S. Sharp, M. Wade, V. (2010) User-Centred Evaluations of Adaptive Systems. Paper Presented at the Doctoral Consortium (Phd Forum), Human Computer Interaction Conference, HCI 2010.
- Murray, T. (2003). An Overview of Intelligent Tutoring System Authoring Tools: Updated Analysis of the State of the art. Authoring Tools for Advanced Technology Learning Environments.
- Ng, M., Hall, W., Maier, P., and Armstrong, R. (2002). The Application and Evaluation of Adaptive Hypermedia Techniques in Web-based Medical Education, *ALT-J*, vol. 10, pp. 19-40.
- Nielsen, J. (1993). *Usability Engineering*, Boston: MA: Academic Press.
- Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, ACM, pp. 152-158.
- Nokelainen, P. (2006). An empirical assessment of pedagogical usability criteria for digital learning material with elementary school students. *Journal of Educational Technology and Society*, vol. 9, pp. 178.
- O'Keeffe, I., Brady, A., Conlan, O., and Wade, V. (2006). Just-in-time Generation of Pedagogically Sound, Context Sensitive Personalized Learning Experiences, *International Journal on E-Learning*, Special Issue on Learning Objects in Context, vol. 5, pp. 113-127.
- Olston, C. and Pandey, S., (2008). Recrawl Scheduling Based on Information Longevity, pp. 437-446.
- Oppermann, R. 1994. Adaptively Supported Adaptability. *International Journal of Human Computer Studies*, vol. 40, pp. 455-472.
- Papert S. (1998). Does Easy Do It? Children, Games, and Learning. *Game Developer Magazine*, vol. 5, pp. 88.
- 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. 383-453.
- Paramythis A. (2009). *Adaptive Systems: Development, Evaluation and Evolution*. PhD, Johannes Kepler Universitat Linz.

- Paramythis A., Weibelzahl S., and Masthoff J. (2010). Layered Evaluation of Interactive Adaptive Systems: Framework and Formative Methods. *User Model User-Adaptation Interaction*, vol. 20, pp 1-71 .
- Pawlowski, J. M., (2003). The European Quality Observatory (EQO): Structuring Quality Approaches for e-Learning, *IEEE*, pp. 209-213.
- Pedhazur, M. J., and Schmelkin, L.P. (1991). *Measurement, Design and Analysis: An Intergrated Approach*, Lawless Erlbaum Associates.
- Peirce, N., Conlan, O., and Wade, V., (2008). Adaptive Educational Games: Providing Non-invasive Personalised Learning Experiences. *Digital Games and Intelligent Toys Based Education*, Second IEEE International Conference, pp. 28-35.
- Peirce, N. and Wade, V. (2010). Personalised Learning for Casual Games: The Language Trap Online Language Learning Game. *Leading Issues in Games Based Learning*, pp. 159.
- Pivec, M. (2007). Editorial: Play and Learn: Potentials of Game-Based Learning. *British Journal of Educational Technology*, vol. 38, pp. 387-393.
- Popescu, E., Trigano, P., and Badica, C. (2007). Towards a Unified Learning Style Model in Adaptive Educational Systems. In *Proceedings of Icalt Conference* , pp. 804-808.
- Raibulet, C. and Masciadri, L., (2009). Evaluation of Dynamic Adaptivity Through Metrics: An Achievable Target?, *IEEE*, pg. 341-344.
- Rieber, L. P. (1996). Seriously Considering Play: Designing Interactive Learning Environments Based on the Blending of Microworlds, Simulations, and Games. *Educational Technology Research and Development*, vol. 44, pp. 43-58.
- Santos Jr, E., Zhao, Q., Nguyen, H., and Wang, H. (2005) Impacts of User Modeling on Personalization of Information Retrieval: An Evaluation with Human Intelligence Analysts.
- Santos, O. C. (2008). A Recommender System to Provide Adaptive and Inclusive Standard-based Support along the e-Learning Life Cycle *Proceedings of the ACM Conference on Recommender Systems*, pp. 319-322
- Scholtz, J. (2004). *Usability Evaluation*. National Institute of Standards and Technology.
- Schulmeister, R. (2004). Instructional Design of University Teaching Point of View-A Plea for Open Learning Situations. *Teaching and New Media, Concepts and Applications in College*, vol. 21, pp. 19-49.

- Lawless, S., O'Connor, A., and Mulwa, C. (2010). A Proposal for the Evaluation of Adaptive Personalised Information Retrieval", Proceedings of the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation Held in Conjunction with ECIR2010 European Conference on Information Retrieval.
- Specht, M., Kravcik, M., Klemke, R., Pesin, L., Huttenhain., (2002). R.: Adaptive Learning Environment for Teaching and Learning in Winds, LNCS, pp. 572-575.
- Tarpin-Bernard, F., Marfisi-Schottman, I. and Habieb-Mammar, H. (2009). AnAmetr: The First Steps to Evaluating Adaptation. In Proceedings of User Modeling, Adaptation and Personalization, UMAP 2009, pp. 11.
- Tintarev, N. and Masthoff, J. (2009). Evaluating Recommender Explanations: Problems Experienced and Lessons Learned for the Evaluation of Adaptive Systems. In Proceedings of User Modeling, Adaptation and Personalization, UMAP 2009.
- Triantafillou, E., Georgiadou, E. and Economides, A. A., (2007). Applying Adaptive Variables in Computerised Adaptive Testing. Australasian Journal of Educational Technology, vol. 23, pp. 350.
- Van Eck, R. (2006). Digital Game-Based Learning: It's Not Just the Digital Natives Who are Restless, Educause Review, vol. 41, pp. 16.
- Van Eck, R. (2007). Building Artificially Intelligent Learning Games. Games and Simulations in Online Learning: Research and Development Frameworks, pp. 271-307.
- Van Velsen, L., Vander Geest, T., Klaasen, R., and Steehouder, M., (2008). User-Centered Evaluation of Adaptive and Adaptable Systems: A Literature Review. The Knowledge Engineering Review, 23, pp. 261-281.
- Verbert, K., Drachler, H., Manouselis, N., Wolpers, M., Vuorikari, R., and Duval, E., (2011). Dataset-driven Research for Improving Recommender Systems for Learning.
- Wade, V. (2009). Challenges for the Multi-Dimensional Personalised Web. In proceedings of User Modeling, Adaptation and Personalization, UMAP 2009.
- Weber, G. and Brusilovsky, P., (2001). Elm-Art: An Adaptive Versatile System for Web-Based Instruction. International Journal of Artificial Intelligence in Education, vol. 12, pp. 351-384.
- Weibelzahl, S. 2003. Evaluation of Adaptive Systems. PhD Thesis, University of Trier.

- Weibelzahl, S. 2005. Problems and Pitfalls in Evaluating Adaptive Systems. In: Fourth Workshop on the Evaluation of Adaptive Systems in Conjunction with UM 2005, pp. 57-66.
- Weibelzahl, S. and Weber, G. (2001). A Database of Empirical Evaluations of Adaptive Systems. pp. 302–306.
- Weibelzahl, S. and Weber, G. (2002). Advantages, Opportunities and Limits of Empirical Evaluations: Evaluating adaptive systems, KI, 16, pp. 17-20.
- Worthen, B. R., Samders, J. R. and Fitzpatrick, J. L. (1997). Program Evaluation. Longman, New York.



## Appendixes

### Appendix A: List of Publications by the Author

#### *Peer- Reviewed (2013-2010)*

- 1 Mulwa, C., and Wade, V. (2013). A Web-Based Evaluation Framework for Supporting Novice and Expert Evaluators of Adaptive E-Learning Systems. *Proceedings of International Conference on E-Technologies and Business on the Web (EBW2013)*, Society of Digital Information and Wireless Communication, pp. 62-67.
- 2 Mulwa, C. Lawless, S., O’Keeffe, I., Sharp, M., and Wade, V. (2012). A Recommender Framework for the Evaluation of End User Experience in Adaptive Technology Enhanced Learning”. *International Journal of Technology Enhanced Learning (IJTL)*, pp. 67-84.
- 3 Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2012). The Evaluation of Adaptive Technology Enhanced Learning Systems. *Proceedings of World Conference on E-Learning in Corporate, Government, and Healthcare and Higher Education (ELEARN)*, Association for the Advancement of Computing in Education (AACE), pp 744-753.
- 4 Mulwa, C., Lawless, Sharp, M. and Wade, V. (2011). The Evaluation of Adaptive and User Adaptive Systems: A Review. *International Journal of Knowledge and Web Intelligence (IJKWI)*, pp 138-156.
- 5 Mulwa, C., Longo, L., Lawless, S., Sharp, M., and Wade, V. (2011). An online framework for supporting the evaluation of personalised information retrieval systems. *Proceedings of the 6th international conference on Ubiquitous and Collaborative Computing*, 2011. British Computer Society. pp 75-85
- 6 Mulwa, C. Lawless, S., Sharp, M., and Wade, V. (2011). A Web based Framework for the Evaluation of End User Experience in Adaptive and Personalised E-Learning Systems. *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE Computer Society, vol. 3, pp 351-356

- 7 Mulwa, C., Lawless, S., Ghorab, M.R., O'Donnell E, Sharp, M., and Wade, V. (2011). A framework for the evaluation of adaptive IR systems through implicit recommendation. In: Andrews, S., Polovina, S., Hill, R., Akhgar, B. (Eds.) *Conceptual Structures for Discovering Knowledge*. Springer Berlin Heidelberg. pp 366-374
- 8 Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems a (Demonstration Paper). *Proceedings of User Modeling, Adaptation and Personalization Conference, (UMAP 2011)*.
- 9 Mulwa, C., Lawless, S., Sharp, M., and Wade, V. (2011). An Evaluation Framework for End User Experience in Adaptive Systems, a (Poster Paper). *Proceedings of User Modeling, Adaptation and Personalization Conference (UMAP 2011)*.
- 10 Mulwa, C., Lawless, S., Sharp, M., Arnedillo-Sanchez, I., and Wade, V. (2010). Adaptive educational hypermedia systems in technology enhanced learning: a literature review. *Proceedings of the 2010 ACM Conference on Information Technology Education*, pp 73-84.
- 11 Mulwa, C., Lawless, S., Sharp, M. and Wade, V. (2010) "OSSES: An Online System for Studies on Evaluation of Systems", *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications*, Association for the Advancement of Computing in Education (AACE), pp 3210-3219.
- 12 Mulwa, C., McDonald, H., O'Keefee, I., Lewis, D., He, Y., and Wade, V. (2012). The Maturity Model: A Novel Way of Evaluating Centre for Next Generation Localisation Demonstrator Systems Based on Industrial Impact, Scientific Collaboration and Interoperability. *Proceedings of World Conference on Educational Media and Technology*. Association for the Advancement of Computing in Education (AACE), pp 1340-1349.
- 13 Mulwa, C., Li, W., Lawless, S., and Jones, G. (2010). A Proposal for the Evaluation of Adaptive Information Retrieval Systems using Simulated Interaction. In *Proceedings of the Workshop on Simulation of Interaction: Automated Evaluation of Interactive IR (SIGIR 2010)*, ACM.
- 14 Mulwa, C., Lawless, S., Sharp, M., Sanchez, I. A., and Wade, V. (2010). Adaptive Educational Hypermedia Systems in Technology Enhanced Learning: A Literature Review. *Proceedings of the 2010 ACM conference on Information Technology Education*, pp 73-84.

- 15 Lawless, S., O'Connor, A. and Mulwa, C. (2010). A Proposal for the Evaluation of Adaptive Personalised Information Retrieval", *Proceedings of the Workshop on Contextual Information Access, Seeking and Retrieval Evaluation held in conjunction with ECIR2010 European Conference on Information Retrieval*.

***Non Peer- Reviewed (2009):***

- 16 Mulwa C., Sharp, M., and Wade, V. User centred Evaluations of Adaptive Systems. Poster presented at the Centre for Next Generation and Localisation (CNGL) Scientific Committee Meeting in October 2009.
- 17 Mulwa. C. Lawless. S. Sharp, M. Wade. V. (2010) User-centred Evaluations of Adaptive Systems. Paper Presented at the Doctoral Consortium (Phd Forum), Human Computer Interaction Conference, .

***Collaboration with other Researchers:***

- 18 O'Donnell, E., Mulwa, C., Sharp, M., and Wade, V. P. (2013). Web-Mediated Education and Training Environments: A Review of Personalised Interactive eLearning. In: *ePedagogy in Online Learning: New Developments in Web Mediated Human Computer Interaction ePedagogy in Online Learning*. Hershey: E. McKay (Ed.). pp. 188-207.
- 19 O' Donnell. E., Mulwa, C., Sharp, M., and Wade. V., (2011). The Human Computer Interaction Issues Associated with the Creation of Personalized Role Playing Simulations. *Irish Human Computer Interaction Conference, Integrated Practice Inclusive Design*. Cork Institute of Technology, pp ---.
- 20 Macarthur, V., Moore, A., Mulwa, C., and Conlan, O. (2011). Towards a Cognitive Model to Support Self-Reflection, Emulating Traits and Tasks in Higher-Order Schemata. *Sixth European Conference on Technology Enhanced Learning*. LNCS

## Appendix B: Implementation of EFEx

### B1. Glossary of Ingredients of the Recommender Component

Glossary Ingredients of the Recommendation	
<b>Glossary</b>	<p>P: Publication (study)            S: System            V: Variation Type            Evp: Evaluation Purpose            N: Number of ...            T: Total number of ...</p>
<b>Factors (cases)</b>	
1 <sup>st</sup> Factor:	<p>Np: Number of publications (study) published that used the evaluation approach.            Tp: Total number of publications in the database</p>
2 <sup>nd</sup> Factor	<p>Tj: Total number of Journal papers in the publications database table            Tc: Total number of Conference papers in the publications database table            Tw: Total number of Workshop papers in the publications database table            Tv: Total venue score of all publications. (calculated as: <math>4 * Tj + 2 * Tc + 1 * Tw</math>)            Nj: Number of Journal papers in which the evaluation approach used was published            Nc: Number of Conference papers in which the evaluation approach used was published.            Nw: Number of Workshop papers which the evaluation approach used was published.            Nv: Venue score of the evaluation approach. (calculated as, <math>4 * Nj + 2 * Nc + 1 * Nw</math>)</p>
3 <sup>rd</sup> Factor	<p>Tsv: Total number of systems that belong to a variation type (v).            Nsv: Number of systems that belong to the variation type (category) that were evaluated using the evaluation approach.</p>
4 <sup>th</sup> Factor	<p>Tep: Total number of evaluation purposes selected by the user (i.e. how many check boxes he/she checked on the screen).            Nevp: Number of selected evaluation purposes that are associated with the evaluation approach.</p>

## B2. Factors Considered When Recommending a Method

<b>Glossary of the Factors considered when recommending a method</b>	
<b>Glossary</b>	
P:	Publication (published study)
S:	System
V:	Variation Type (category)
E <sub>VP</sub> :	Evaluation Purpose (Goal)
N:	Number of ...
T:	Total Number of...
	<b>1<sup>st</sup> Factors</b>
N <sub>p</sub> :	Number of publications published that used the evaluation method.
T <sub>p</sub> :	Total number of publications in the database
	<b>2<sup>nd</sup> Factors</b>
T <sub>j</sub> :	Total number of Journal Papers in the publications table
T <sub>C</sub> :	Total number of Conference Papers in the publication table
T <sub>w</sub> :	Total number of workshop papers in the publication table.
T <sub>v</sub> :	Total venue score of all publications (i.e. calculated as: $4 * T_j + 2 * T_c + 1 * T_w$ ).
N <sub>j</sub> :	Number of Journal Papers in which the evaluation method was used.
N <sub>C</sub> :	Number of Conference papers in which the evaluation method was used.
N <sub>w</sub> :	Number of Workshop papers in which the evaluation method was used
N <sub>v</sub> :	Venue scor of the evaluation method (calculated as $4 * N_j + 2 * N_c + 1 * N_w$ )
	<b>3<sup>rd</sup> Factors</b>
T <sub>SV</sub> :	Total number of systems that belong to a given variation type (v)
N <sub>SV</sub> :	Number of systems that belong to the given variation type (v) that were evaluated using the evaluation method.
	<b>4<sup>th</sup> Factors</b>
T <sub>EVP</sub>	Total number of evaluation purposes (goal) selected by the user (i.e. how many checkboxes the user checked on the screen).
N <sub>EVP</sub> :	Number of selected evaluation purposes that are associated with the evaluation method.

### B3. Implementing a bundle (Method, Criteria & Metric)

*Java Code*

**Recommended bundles for a list of given evaluation methods**

```
package effee.formbean;
import java.util.*;
import java.sql.*;

/**
 * This is NOT a bean. This is just a class to hold all the work
 * that we're doing to create recommended bundles for a list of given
 * methods.
 * @author admin
 */
//This class also does the explanations on how we arrive (get) the
recommended bundle.
public class BundlingWork
{
    public static Vector<BundleItem>
    recommendBundles(Hashtable<String, Double>
    methodsAndScoresTable)
    {
        //System.out.println("I am in the bundling method.");
        Vector<BundleItem> recommendedBundleItemsVector = new
        Vector<BundleItem>(100);
        try
        {
            //1) Connect to Database and prepare statements

            DriverManager.registerDriver(new com.mysql.jdbc.Driver());
            Connection con =
            DriverManager.getConnection("jdbc:mysql://localhost/effee",
            "OSSES", "password");
            PreparedStatement
            getMethodIdStatement=con.prepareStatement("select id from
            evaluation_methods_adaptive_systems where name=?");
            PreparedStatement
            getMostOccuringCriteriaIdStatement=con.prepareStatement("select
            criteria_id, count(*) from evaluated_adaptive_previoushistory
            where method_id=? group by criteria_id order by count(*)
            desc");
            PreparedStatement
            getMostOccuringMetricIdStatement=con.prepareStatement("select
            metric_id, count(*) from evaluated_adaptive_previoushistory
            where method_id=? AND criteria_id=? group by metric_id order by
            count(*) desc");
            PreparedStatement
            getMethodNameStatement=con.prepareStatement("select name from
            evaluation_methods_adaptive_systems where id=?");
            PreparedStatement
            getCriteriaNameStatement=con.prepareStatement("select name from
            evaluation_criteria_adaptive_systems where id=?");
            PreparedStatement
            getMetricNameStatement=con.prepareStatement("select name from
            evaluation_metrics_adaptive_systems where id=?");

            Enumeration<String> methods = methodsAndScoresTable.keys();
            String name = null;
            String[] methodNames = new
            String[methodsAndScoresTable.size()];
```

```

//int[] methodScores = new int[methodsAndScoresTable.size()];
double[] methodScores = new
double[methodsAndScoresTable.size()];
    int n = 0;
    while(methods.hasMoreElements())
    {
        name = methods.nextElement();
        methodNames[n] = name;
        //Integer iobj = methodsAndScoresTable.get(name);
        Double iobj = methodsAndScoresTable.get(name);
        methodScores[n] = iobj;

        n++;
    }
//sort here.
boolean sorted = false;
int j = 0;
String tempName = null;
//int tempScore = 0;
double tempScore = 0.0;
while (!sorted)
{
    sorted = true;
    j++;
    for (int k=0; k<methodScores.length - j; k++)
    {
        if (methodScores[k] <methodScores[k+1])
        {
            tempScore = methodScores[k];
            tempName = methodNames[k];

            methodScores[k] = methodScores[k+1];
            methodNames[k] = methodNames[k+1];

            methodScores[k+1] = tempScore;
            methodNames[k+1] = tempName;

            sorted = false;
        }
    }
}
String description = null;
for(int i=0 ; i<methodNames.length; i++)
{
    description = "This bundle is recommended because of the
following:";
//2) Execute a query to give us the id of the method using the name of
the method that we have.
    getMethodIdStatement.setString(1,methodNames[i]);
    ResultSet rs1 =
    getMethodIdStatement.executeQuery();
    int methodId = 0;
    if(rs1.next())
    {
        methodId=rs1.getInt(1);
    }
    rs1.close();

    if(i==0)
    {
        description += " The method " + methodNames[i] + " is the
most recommended method.";
    }
    else
    {

```

```

        description += " The method " + methodNames[i] + " is among
        the recommended methods.";
    }
//3.1) execute the prepared query that retrieves the criteria id that
has the maximum count for the method at hand from the previous history
table

    getMostOccuringCriteriaIdStatement.setInt(1,methodId);
        ResultSet rs2 =
    getMostOccuringCriteriaIdStatement.executeQuery();
        int criteriaId = 0;
        int criteriaCount = 0;
        if(rs2.next())
        {
            criteriaId=rs2.getInt(1);
            criteriaCount = rs2.getInt(2);
        }
        rs2.close();

//3.2) execute the prepared query that retrieves the metric id that
has the maximum count for the method and criteria obtained from
previous step

    getMostOccuringMetricIdStatement.setInt(1,methodId);

    getMostOccuringMetricIdStatement.setInt(2,criteriaId);
        ResultSet rs3 =
    getMostOccuringMetricIdStatement.executeQuery();
        int metricId = 0;
        int metricCount = 0;
        if(rs3.next())
        {
            metricId=rs3.getInt(1);
            metricCount = rs3.getInt(2);
        }
        rs3.close();
//3.3) Now, that I have the bundle elements: method_id, criteria_id,
metric_id
// execute three queries to get the names of the ids I have for the
method, the criterion, and the metric

        getMethodNameStatement.setInt(1,methodId);
            ResultSet rs4 =
        getMethodNameStatement.executeQuery();
        String methodName = null;
        if(rs4.next())
        {
            methodName=rs4.getString(1);
        }
        rs4.close();
        getCriteriaNameStatement.setInt(1,criteriaId);
            ResultSet rs5 =
        getCriteriaNameStatement.executeQuery();
        String criteriaName = null;
        if(rs5.next())
        {
            criteriaName=rs5.getString(1);
        }
        rs5.close();

        getMetricNameStatement.setInt(1,metricId);
            ResultSet rs6 =
        getMetricNameStatement.executeQuery();
        String metricName = null;

```



```

        if(rs6.next())
        {
            metricName=rs6.getString(1);
        }
        rs6.close();
                if(criteriaName == null ||
                criteriaName.trim().equals(""))
        {
                if(metricName == null ||
                metricName.trim().equals(""))
        {
                description+= " At the moment
                there are no criteria or metrics that are specifically
                recommended to be used with this method.";
            }
            else
            {
                description+= " At the moment
                there are no criteria that are specifically recommended
                to be used with this method.";
                description+= " The metric
                "+metricName+" was the highest occurring metric associated with
                the method and the criterion in the reference database (used
                "+metricCount+" times with the method and the criterion).";
            }
        }
        else
        {
            description+= " The criterion
            "+criteriaName+" was the highest occurring criterion associated
            with the method in the reference database (used
            "+criteriaCount+" times with the method).";
            if(metricName == null ||
            metricName.trim().equals(""))
            {
                description+= " At the moment there are
                no metrics that are specifically recommended to be used with
                this method.";
            }
            else
            {
                description+= " The metric
                "+metricName+" was the highest occurring metric associated with
                the method and the criterion in the reference database (used
                "+metricCount+" times with the method and the criterion).";
            }
        }
        //3.4) Create a new BundleItem object with the
        three names (method name, criteria name, metric name)
        BundleItem item = new
        BundleItem(methodName,criteriaName,metricName, description,
        methodScores[i]);
//3.5) Add the object of the BundleItem to the vector of bundles
        recommendedBundleItemsVector.add(item);
    }
}
catch(SQLException ex)
{
    ex.printStackTrace();
}
//4) return the vector
/*
//teporary test to print the vector
System.out.println("I am going to print the bundles

```

```

now:");
    for(int i=0 ; i<recommendedBundleItemsVector.size(); i++)
    {
        BundleItem bi = recommendedBundleItemsVector.get(i);
        System.out.println("Bundle#"+(i+1));
        System.out.println("Method Name: "+ bi.getMethod());
        System.out.println("Criteria Name: "+
        bi.getCriteria());
        System.out.println("Metric Name: "+ bi.getMetric());
    }
    */
    return recommendedBundleItemsVector;
}
}

```

## B4. Calculating Total Score for Each Metric

### Calculating the Total Score for Each Metric and Then Sorting in Descending Order the Calculations Produced

```

//Finally, calculate the total score for each metric then sort descending.

double[] finalScoresForMetrics = new double[numberOfMetrics];
for(int i=0; i<numberOfMetrics; i++)
{
    finalScoresForMetrics[i] = normalizedFactor1ScoresForMetrics[i] +
    normalizedFactor2ScoresForMetrics[i]
        + normalizedFactor3ScoresForMetrics[i] ;
}
int originalIndexOfHighestMetric = 0;
double maxFinalScoreForMetrics = finalScoresForMetrics[0];
for(int i=1; i<finalScoresForMetrics.length; i++)
{
    if(finalScoresForMetrics[i]>maxFinalScoreForMetrics)
    {
        maxFinalScoreForMetrics = finalScoresForMetrics[i];
        originalIndexOfHighestMetric = i;
    }
}
currentEvaluation.setRecommendedMetricsScores(finalScoresForMetrics);

int[] sortedMetricsIds = metricsIds.clone();
String[] sortedMetricsNames = metricsNames.clone();
double[] sortedFinalScoresForMetrics = finalScoresForMetrics.clone();

boolean sorted = false;
int j = 0;
String tempName = null;
int tempId = 0;
double tempScore = 0.0;
while (!sorted)
{
    sorted = true;
    j++;
    for (int k=0; k<sortedFinalScoresForMetrics.Length - j;
    k++)
    {

```

```

        if (sortedFinalScoresForMetrics[k]
        <sortedFinalScoresForMetrics[k+1])
        {
            tempScore = sortedFinalScoresForMetrics[k];
            tempName = sortedMetricsNames[k];
            tempId = sortedMetricsIds[k];

            sortedFinalScoresForMetrics[k] =

            sortedFinalScoresForMetrics[k+1];

            sortedMetricsNames[k] =
            sortedMetricsNames[k+1];
            sortedMetricsIds[k] = sortedMetricsIds[k+1];

            sortedFinalScoresForMetrics[k+1] = tempScore;
            sortedMetricsNames[k+1] = tempName;
            sortedMetricsIds[k+1] = tempId;

            sorted = false;
        }
    }
}
currentEvaluation.setSortedDescendinglyRecommendedMetricsIds (sortedMetrics
Ids);
currentEvaluation.setSortedDescendinglyRecommendedMetricsNames (sortedMetri
csNames);
currentEvaluation.setSortedDescendinglyRecommendedMetricsScores (sortedFina
lScoresForMetrics);

    for(int i=0; i<sortedMetricsNames.length && i<5; i++)
    {

```

## B5. EFX Technical Design

### Technical Design

The technical design includes several well-integrated technologies and the functionalities of all the components discussed in Section 4.3.1. The technologies used include: NetBeans 6.8 platform<sup>37</sup>, Apache Lucene, Apache\_OpenJPA, Apache-Tomcat 5.5,<sup>38</sup> Myfaces-core, MySql-win32, MySql-connector-java, Json<sup>39</sup> and Google Translate<sup>40</sup> and Html. JavaServer Pages (JSP) is a technology that dynamically generates web pages based on html, xml and other document types.

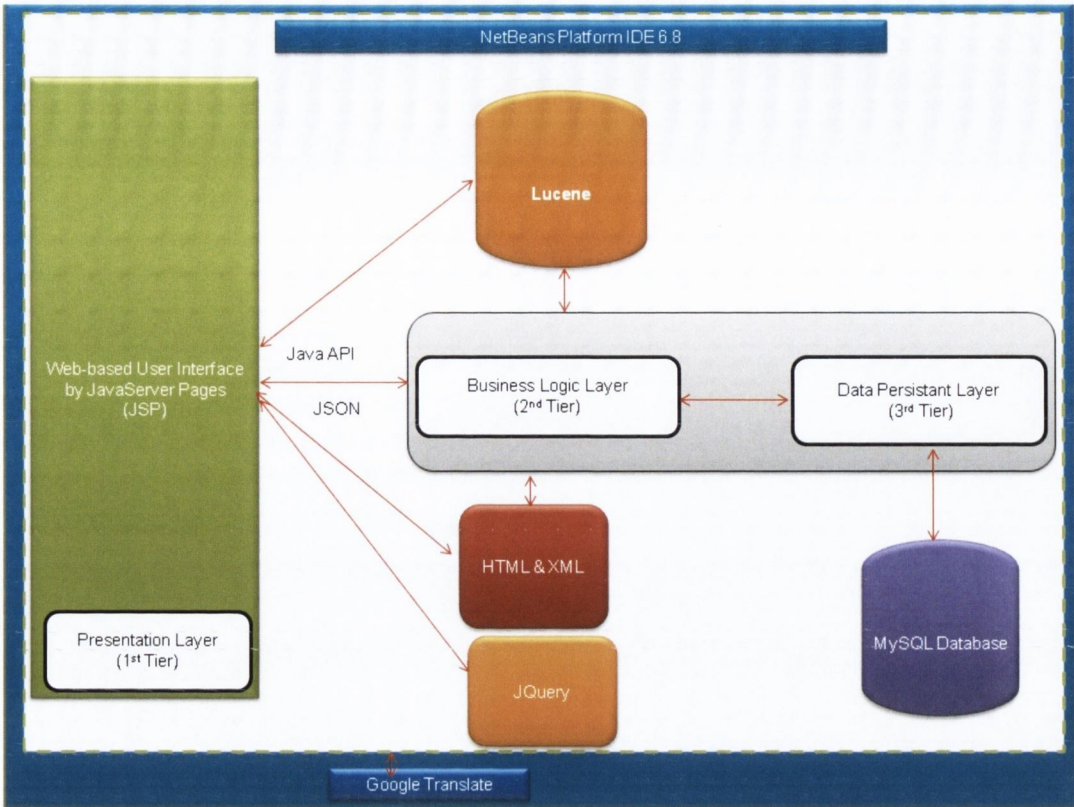
---

<sup>37</sup> <https://netbeans.org/>

<sup>38</sup> <http://tomcat.apache.org/>

<sup>39</sup> <http://www.json.org/>

<sup>40</sup> <https://developers.google.com/translate/>



## Appendix C: Eliciting Knowledge-base for EFEx

### *Evaluations of Adaptive Systems Developed from 2000 to 2012*

**Q1.** Have You Developed an Adaptive System in the Past (from 2000 to 2011)? (i.e. An adaptive system refers to a system which tailors its output, using Implicit Inferences based on interaction with the user)

- a. Yes                      b. No

**Q2.** If You Answered Yes to Q1, Please Provide:

- a. Name of Adaptive System Developed,
- b. Year the System was Developed,
- c. Other Details:

**Q3.** If You Have Developed an Adaptive System(S), what was improved by Adaptivity?

**Q4.** What is the Variation Type of the Adaptive System You have Developed?

- Adaptive Educational Hypermedia System
- Adaptive Information Retrieval System
- Online Help Customer Care System

Other Variation Types (please specify)

**Q5. Please Tick the Meta Data Models Your System Uses**

- |  |   |   |
|--|---|---|
| <input type="checkbox"/> User model    | <input type="checkbox"/> Presentation model | <input type="checkbox"/> Task model       |
| <input type="checkbox"/> Domain model  | <input type="checkbox"/> Device model       | <input type="checkbox"/> Strategy model   |
| <input type="checkbox"/> Content model | <input type="checkbox"/> System model       | <input type="checkbox"/> Navigation model |

Other models:

**Q6. If You Conducted a Whole-System Evaluation, What Evaluation Methods did you use?**

- |   |  |  |
|---|--|--|
| <input type="checkbox"/> Interviews                 | <input type="checkbox"/> Cross-Validation        | <input type="checkbox"/> Experimental Evaluation         |
| <input type="checkbox"/> Questionnaires             | <input type="checkbox"/> Heuristic-E valuations  | <input type="checkbox"/> Cultural Probes                 |
| <input type="checkbox"/> Focus Group                | <input type="checkbox"/> Prototyping             | <input type="checkbox"/> User Test                       |
| <input type="checkbox"/> Discussion Group           | <input type="checkbox"/> Expert Review           | <input type="checkbox"/> Creative Brainstorming Sessions |
| <input type="checkbox"/> User Observation           | <input type="checkbox"/> Parallel Design         | <input type="checkbox"/> Empirical Observations          |
| <input type="checkbox"/> The Systematic Observation | <input type="checkbox"/> Cognitive Walkthroughs  | <input type="checkbox"/> Ethnographic Observa            |
| <input type="checkbox"/> Verbal Protocol            | <input type="checkbox"/> Social-technical Models | <input type="checkbox"/> Quantative                      |
| <input type="checkbox"/> Data Mining                | <input type="checkbox"/> Wizard of Oz Simulation | <input type="checkbox"/> Grounded Theory                 |
| <input type="checkbox"/> Play With Layer            | <input type="checkbox"/> Scenario Based Design   | <input type="checkbox"/> Cooperative Evaluation          |
| <input type="checkbox"/> Simulated Users            | <input type="checkbox"/> Usability Testing       |  |

Other Evaluation Methods Used

**Q7.** If you conducted a whole evaluation, what criteria did you use?

- |   |  |   |
|---|--|---|
| <input type="checkbox"/> Usability                            | <input type="checkbox"/> Usability of Interface Adaptation | <input type="checkbox"/> User Performance   |
| <input type="checkbox"/> Perceived Usefulness                 | <input type="checkbox"/> User Satisfaction                 | <input type="checkbox"/> Transparency   |
| <input type="checkbox"/> Intention to Use                     | <input type="checkbox"/> Interface Knowledge               | <input type="checkbox"/> Appropriateness  |
| <input type="checkbox"/> Trust and Privacy Issues             | <input type="checkbox"/> Hyperspace Experience Workload    | <input type="checkbox"/> User Cognitive   |
| <input type="checkbox"/> Appropriateness of Adaptation        | <input type="checkbox"/> Early Prototype Evaluations       | <input type="checkbox"/> Real User Actions  |
| <input type="checkbox"/> User Behaviour                       | <input type="checkbox"/> Evaluation before Implementation  | <input type="checkbox"/> To Combine Qualitative Evaluation                          |
| <input type="checkbox"/> User Goal                            | <input type="checkbox"/> Content Adaptation                | <input type="checkbox"/> To Discover New Theories                                   |
| <input type="checkbox"/> Knowledge of Domain                  | <input type="checkbox"/> Preferences                       | <input type="checkbox"/> Evaluation of Vertical or Horizontal Prototype             |
| <input type="checkbox"/> Background and Hyperspace Experience | <input type="checkbox"/> User Skills and Capabilities      | <input type="checkbox"/> Collaboration with real users during final evaluation step |

Other Evaluation Criteria Used (please specify)

**Q8.** If You Conducted Evaluations of Specific Metadata Models of Adaptive System, What Evaluation Methods did you use?

(For each model evaluated, please indicate which evaluation methods and criteria you used)

**Q9.** During this Evaluation (Conducted in Question 6 and 7 above), What Metrics did You Use to Measure Performance against these criteria?

- |   |   |   |
|---|---|---|
| <input type="checkbox"/> Accuracy of Recommendations                      | <input type="checkbox"/> Software Size and Length Metrics         | <input type="checkbox"/> AvgpACF: Average Personalisation Adaptive Cost Per Functionality |
| <input type="checkbox"/> Accuracy of Retrieval                            | <input type="checkbox"/> UiAI: User Interaction                   | <input type="checkbox"/> MpOCF: Minimum Personalisation Overall Cost                      |
| <input type="checkbox"/> AiAI: Administrator Interaction Adaptivity Index | <input type="checkbox"/> pLatency: Performance Latency            | <input type="checkbox"/> pOCF: Personalisation Overall Cost per Functionality             |
| <input type="checkbox"/> Behavioural Complexity                           | <input type="checkbox"/> pQoR: Performance Quality on Response    | <input type="checkbox"/> ApOC: Adaptive Personalisation Overall Cost                      |
| <input type="checkbox"/> Reliability Metrics                              | <input type="checkbox"/> pIA: Performance Influence on Adaptivity | <input type="checkbox"/> DSAI: Domain Specific Adaptivity Index                           |
| <input type="checkbox"/> Precision  |   |   |
| Personalisation Adaptive  |   |   |

Other Metrics (please specify)

## Appendix D: Evaluation Framework for End User Experience in Adaptive Systems (EFEx)

### D1. Post-Questionnaire: Design Testing of EFEx

Potential Benefits of EFEx Framework:

- i. Repository for User-Centred and Layered Evaluations of Adaptive Systems.
- ii. Recommendations on how to evaluate an existing adaptive system or a new adaptive system.
- iii. Recommendations on how to evaluate metadata models of adaptive systems.
- iv. User-centred evaluation methodology for adaptive systems.
- v. Taxonomy of technical terms of evaluations adaptive systems.
- vi. A Translate that translates information into users' language of choice

**Q1.** Which of the following features of EFEx Framework would you find (consider) useful?

- Recommendations on how to evaluate an adaptive system/metadata models of these systems/Authoring Adaptive tools (i.e. how to combine and apply existing evaluation methods (techniques), metrics and measurement criteria in order to evaluate the adaptive system and the metadata models (i.e. user, domain, strategy, task, content, device, system, navigation and presentation models) used by this system).
- A centralised repository which stores (i.e. layered, UCE, metadata models and authoring adaptive technologies) evaluation studies from 2000 to date
- A methodology which illustrates or explains how to apply user-centred evaluation techniques.
- A Taxonomy of technical terms of evaluation of Adaptive Systems
- A Translate that translates information into users' language of choice
- ALL

Other Features You Recommend (please specify)



## **D2. Experiment 1 - Hybrid (Case-based & Knowledge-based) Recommender System**

### **D2.1 Expert Evaluators: Task-based Evaluation Experiment (Recommending Evaluation Techniques)**

To participate in this study, you must be familiar with adaptive systems or the evaluation such systems. This task is will take you 30 minutes to complete.

#### **Task 1:**

You will be asked to choose one category (variation type) of adaptive systems which you would wish to focus on during evaluation (i.e. Adaptive Educational Hypermedia System, Adaptive Information Retrieval System, online help customer care system, adaptive learning system, adaptive recommender system, intelligent tutoring systems and adaptive public displays). You will be asked to choose different evaluation techniques (i.e. approaches, methods, measurement criteria and metrics) you would recommend when evaluating systems belonging to the category selected.

#### **Task 2: (section D3.2)**

In this part you will be interacting with an automated hybrid (case-based and knowledge-based) recommender system. This system will recommend to you which evaluation approach and techniques (methods, metrics and criteria) to use when evaluating an adaptive system. The system will also bundle the recommended techniques into the most appropriate method to use together with measurement criteria(s) and metric. Throughout the process of recommendations, you will be provided with explanations as to how the recommended techniques were derived. To interact with the recommender system, I will need to provide you with login details and a link to the recommender system. In this regard we require an email address so that we can set up the login and email the link to the system to you.

**Q1.** How would you rate your evaluation skills of adaptive systems

- a.  Very experienced evaluator (3+ yrs Research experience)
- b.  Experienced evaluator (1-3 years) research experience)
- c.  I find most evaluations of adaptive system difficult to understand
- d.  I have no skills in evaluations of such systems
- e.  I would like to learn how to evaluate adaptive systems.

If you answered:

- c. I find most evaluations of adaptive system difficult to understand,
- d. I have no skills in evaluations of such systems and
- e. I would like to learn how to evaluate adaptive systems. Task 1 is not suitable for you, but

### ***D2.1.1 Recommending Evaluation Approaches & Techniques for Existing adaptive Systems***

#### ***Recommending an Evaluation Technique:***

In this task, you are required to choose one variation type (category) of adaptive system which you wish to focus on during this study. You will be asked to choose which properties of the adaptive system (i.e. belonging to the variation type you have chosen) you would feel comfortable recommending evaluation technique(s) for. Next you will be asked which evaluation technique(s) you would recommend to evaluate such an adaptive system(s).

**Q2.** Please choose which category (i.e. also known as variation type) of adaptive systems you wish to focus on during evaluation (\*Please choose one category only).

(\* We have provided examples of adaptive systems belonging to the different variation types)

- |   |  |   |
|---|--|---|
| <input type="radio"/> Adaptive Hypermedia -<br>[e.g. KBS Hyperbook<br>(2000), APeLS (2002), MOT<br>(2003) ] | <input type="radio"/> Adaptive Recommender<br>System<br>[ e.g. ExDis (2011), News<br>semantic recommender<br>system (2009) ] | <input type="radio"/> General User Interfaces [<br>e.g. (Mouse Smoothing<br>Algorithms for Users<br>with Tremors (2008) ] |
|---|--|---|

- Adaptive Educational Hypermedia System [e.g. Activemath (2000-ongoing), English-Math ABLE (2007 2008), aLFanet (2005), English ABLE (2006)]
- Adaptive Public Displays [ e.g. Contextual Display (2010), Friend Finder, Late-o-Meter ]
- Museum visitors guide [e.g. PEACH (2004), PIL (2007), APIL (2010)]
- General User Interfaces
- Adaptive Information Retrieval System [e.g. MovieLens (2006), SuggestBot (2005), Bifrost (2010)]
- Intelligent Tutoring Systems [e.g. ERM-Tutor (2005), EER- Tutor (2003), Thermo-Tutor (2010),]
- General Purpose Adaptive Systems [e.g. AHA! (1996-2007), GALE 2011)]
- Online Help Customer Care System [ e.g. CID (2008), Viper (2008), PIA, Adaptive Information Retrieval and Composition System (2010) ]
- Adaptive Learning System [e.g. i OLMlets (2006)]
- Adaptive Web Based Systems [e.g. UNITE (2011), www.assistent.org Developer 2002 ]
- Adaptive Educational Game [ e.g. Adaptive Educational Game (2010) ]

Properties of adaptive systems you feel comfortable recommending an evaluation technique:

In Questions 3 to 7 you are required to choose which property(s), you would feel comfortable recommending an evaluation technique (i.e. method, criteria and metric) to evaluate an adaptive system (i.e. belonging to the variation type (category) you choose in Question Q2.

**Q3.** Please select which system characteristics you wish to focus on during evaluation.

(\* Select as many as you like)

- Adaptable Functionality and Appearance
- Adaptable Interfaces
- Sensitivity (i.e. the degree to be affected by or responsive to some environmental stimuli, i.e. the impact potential of the system).
- Susceptibility (i.e. the extent to which a system is open, liable, or sensitive to
- Responsiveness (i.e. responsiveness of an adaptive system is the magnitude or degree to react to stimuli.
- Stability (i.e. in this case describes to which extent a system is not easily moved or modified from a stable state.
- Feedback (i.e. the circular causality of feedback loops is taken into account for

environmental stimuli.

Robustness (i.e. as in a general context, is the strength or degree to which a system is not given to influence).

Adaptive capacity (i.e. the potential or capability to adapt something according to environmental stimuli or their effects)

Adaptability (i.e. the property of being adaptable. If something adapts itself dynamically, then it is said to be adaptive and adaptable.

Cognitive

History

Other (please specify other system characteristics you wish to focus on)

regulation processes, i.e. for ensuring the success of system's efforts to maintain equilibrium or to reach a goal. Further, integral feedback control is not only sufficient but also necessary for robust adaptation.

Fitness (i.e. a measurable degree that depends on the reproductive success of adaptations. Equivalently to fitness, sometimes the term efficiency is used as a synonym.

Predictability (i.e. Predictability comprises the a-priori work in order to pre-compute future behavioural steps, e.g. in order to optimise some results, enhance the fitness or foster the autonomy of the system.

***Evaluation Purpose:***

**Q4.** Based upon the category (variation type) of adaptive system and the system characteristics that you have selected, what would be the goal(s) or purpose(s) of the evaluation being conducted?

(\* Select as many as you like)

Evaluate system effectiveness

Test End User Experience

Test System Performance

Test Adaptivity

Test Usability

Estimate Efficiency of the Instruction

Check that Constructed Metadata Models Accurately Represent Real World

Determine whether the Adaptation Decisions Made are the Optimal Ones

- |   |  |
|---|--|
| <input type="checkbox"/> Check quality of Raw Input Data                | <input type="checkbox"/> Determine whether the Implementation of the   |
| <input type="checkbox"/> Check that Input Data is Interpreted Correctly | Adaptation Decision made is Optimal                                    |
|   | <input type="checkbox"/> Evaluate the overall adaptation theory        |
|   | <input type="checkbox"/> Summative Evaluation of the Adaptation Theory |

Other (please specify other evaluation purpose/goal)

**Q5. What Questions would this Evaluation enable you to answer once it is complete?**

What Kind of question(s) would you wish to answer during evaluation?

(\* Select as many as you like)

- |  |  |
|--|--|
| <input type="checkbox"/> What is Improved by Having Adaptivity?                | <input type="checkbox"/> How to evaluate adaptation as a whole                   |
| <input type="checkbox"/> Why are you testing for end user experience           | <input type="checkbox"/> Questions relating to all layers of the adaptive system |
| <input type="checkbox"/> Questions on modelling the current state of the World |  |

Other (please specify if you have other kind of question(s))

**Q6. Kind of results you would expect to obtain after evaluation:**

If you were to conduct the evaluation in the form that you have selected above, what would it help you to improve in the system?

- |   |  |   |
|---|--|---|
| <input type="checkbox"/> End user experience  | <input type="checkbox"/> Accuracy of recommendations       | <input type="checkbox"/> Time taken to perform a task |
| <input type="checkbox"/> System Performance   | <input type="checkbox"/> System Efficiency                 | <input type="checkbox"/> Learning experience          |
| <input type="checkbox"/> System Effectiveness | <input type="checkbox"/> Accuracy of information retrieved | <input type="checkbox"/> Quality of raw input data    |

Other (Please specify if you have other kind of results)

Recommending an Evaluation Approach:

From Question 8 to 10, we will ask you which evaluation technique(s) you would feel comfortable recommending when evaluating an adaptive system belonging to the variation type you choose in Question 3.

To determine which evaluation approach(s) and techniques to recommend please use the characteristics you selected. Please rank your recommendations from 5 to 1. (i.e. 5 being the most appropriate technique and 1 the least)

**Q7.** Which of the following evaluation approach (s) would you recommend to be used when evaluating an adaptive system(s) belonging to the variation type (i.e. you choose in Q2.)?

Please rate the recommended approach

	<b>[5] Most Appropriate</b>	<b>[4] Appropriate</b>	<b>[3] Somehow Appropriate</b>	<b>[2] Least appropriate</b>	<b>[1] Not appropriate</b>
User - Centered Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Layered Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utility - Based Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Empirical Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Heuristic Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Recommending Evaluation Methods, Criteria and Metrics:

**Q8.** Which of the following evaluation methods do you feel are appropriate for evaluating systems of the variation type you choose in Q3?

Please use the drop-down lists below. You can recommend one or more evaluation methods by rating them from most appropriate to not appropriate.

	<b>Evaluation Methods (Recommend 1)</b>	<b>Evaluation Methods (Recommend 2)</b>	<b>Evaluation Methods (Recommend 3)</b>
Most Appropriate	<div style="border: 1px solid black; padding: 2px;">           Data Mining           <div style="float: right;">▼</div> </div>	<div style="border: 1px solid black; padding: 2px; height: 15px;"> <div style="float: right;">▼</div> </div>	<div style="border: 1px solid black; padding: 2px; height: 15px;"> <div style="float: right;">▼</div> </div>
Appropriate	<div style="border: 1px solid black; padding: 2px;">           Questionnaires            Interviews            Usability Testing            Experimental Evaluation            Focus Group            User Observation            User Test            Expert Review            Quantative  <span style="background-color: #e0e0e0;">Data Mining</span>            Empirical Observations            Simulated Users            Cross-Validation            Heuristic Evaluations            Wizard of Oz Simulation            Creative Brainstorming Sessions            Eye-Tracking            Task Completion Time            System Preference Survey            Formative Evaluation            Summative Evaluation         </div>	<div style="border: 1px solid black; padding: 2px; height: 15px;"> <div style="float: right;">▼</div> </div>	<div style="border: 1px solid black; padding: 2px; height: 15px;"> <div style="float: right;">▼</div> </div>
Somehow Appropriate		<div style="border: 1px solid black; padding: 2px; height: 15px;"> <div style="float: right;">▼</div> </div>	<div style="border: 1px solid black; padding: 2px; height: 15px;"> <div style="float: right;">▼</div> </div>

**Q9.** Which of the following evaluation criteria do you feel are appropriate for evaluating systems of the variation type you choose in Q2?

Please use the drop-down lists below. You can recommend one or more evaluation criteria by rating them from most appropriate to not appropriate.

	Measurement Criteria (Recommend 1)	Measurement Criteria (Recommend 2)	Measurement Criteria (Recommend 3)
Most Appropriate	<div style="border: 1px solid black; padding: 2px;"> <input type="text" value=""/> <ul style="list-style-type: none"> <li>User Satisfaction</li> <li>User Performance</li> <li>Usability</li> <li>Perceived Usefulness</li> <li>User Skills and Capabilities</li> <li>Intention to Use</li> <li>Preferences</li> <li>Trust and Privacy Issues</li> <li>Content Adaptation</li> <li>Appropriateness of Adaptation</li> <li>User Behaviour</li> <li>Knowledge of Domain</li> <li>User Goal</li> <li>Usability of Interface Adaptation</li> <li>Interface Knowledge</li> <li>Precision</li> <li>Recall</li> <li>Early Prototype Evaluations</li> <li>Evaluation before Implementation</li> <li>Transparency</li> <li>Effectiveness (for decision support)</li> <li>Appropriateness</li> <li>User Cognitive Workload</li> <li>Real User Actions</li> <li>Piloting</li> <li>System Response Time</li> <li>To Combine Qualitative Evaluation</li> <li>Collaboration with Real Users during Final Evaluation</li> </ul> </div>	<input type="text" value=""/>	<input type="text" value=""/>
Appropriate		<input type="text" value=""/>	<input type="text" value=""/>
Somehow Appropriate		<input type="text" value=""/>	<input type="text" value=""/>

**Q10.** Which of the following evaluation metrics do you feel are appropriate for evaluating systems of the variation type you choose in Q2? Please use the drop-down lists below. You can recommend one or more evaluation metrics by rating them from most appropriate to not appropriate.

	Measurement Criteria (Recommend 1)	Measurement Criteria (Recommend 2)	Measurement Criteria (Recommend 3)
Most Appropriate	<div style="border: 1px solid black; padding: 2px;"> <input type="text" value="Presence Questionnaire Scores"/> <ul style="list-style-type: none"> <li>Accuracy of Recommendations</li> <li>Precision</li> <li>Recall</li> <li>Accuracy of Retrieval</li> <li>Reliability Metrics</li> <li>Task Completion Time</li> <li>Task Effectiveness</li> <li>Task Efficiency</li> <li>Perceived Appropriateness of Adaptations</li> <li>Invasiveness of Adaptations</li> <li>Awareness of Adaptations</li> <li>Behavioural Complexity</li> <li>Accuracy of Models</li> <li>pIA: Performance Influence on Adaptivity</li> <li>ApOC: Adaptive Personalisation Overall Cost</li> <li>Students' Content Learning Gains</li> <li>Presence Questionnaire Scores</li> <li>Intrinsic Motivation Inventory Scores</li> </ul> </div>	<input type="text" value=""/>	<input type="text" value=""/>
Appropriate		<input type="text" value=""/>	<input type="text" value=""/>
Somehow Appropriate		<input type="text" value=""/>	<input type="text" value=""/>



11. Which of the following evaluation techniques (i.e. the techniques you recommended in Q8-Q10) would you bundle to be used together?

The term "bundle" refers to appropriate combination of a method/criteria/metric that can be used together when evaluating the properties you selected in Q3-Q6.

	Evaluation Methods	Measurement Criteria	Evaluation Metric
Most Appropriate Bundle (Bundle 1)	<ul style="list-style-type: none"> <li>Data Mining</li> <li>Questionnaires</li> <li>Interviews</li> <li>Usability Testing</li> <li>Experimental Evaluation</li> <li>Focus Group</li> <li>User Observation</li> <li>User Test</li> <li>Expert Review</li> <li>Quantitative</li> <li>Data Mining</li> <li>Empirical Observations</li> <li>Simulated Users</li> <li>Cross-Validation</li> <li>Heuristic Evaluations</li> <li>Wizard of Oz Simulation</li> <li>Creative Brainstorming Sessions</li> <li>Eye-Tracking</li> <li>Task Completion Time</li> <li>System Preference Survey</li> <li>Formative Evaluation</li> <li>Summative Evaluation</li> </ul>	<ul style="list-style-type: none"> <li>User Satisfaction</li> <li>User Performance</li> <li>Usability</li> <li>Perceived Usefulness</li> <li>User Skills and Capabilities</li> <li>Intention to Use</li> <li>Preferences</li> <li>Trust and Privacy Issues</li> <li>Content Adaptation</li> <li>Appropriateness of Adaptation</li> <li>User Behaviour</li> <li>Knowledge of Domain</li> <li>User Goal</li> <li>Usability of Interface Adaptation</li> <li>Interface Knowledge</li> <li>Precision</li> <li>Recall</li> <li>Early Prototype Evaluations</li> <li>Evaluation before Implementation</li> <li>Transparency</li> <li>Effectiveness (for decision support)</li> <li>Appropriateness</li> <li>User Cognitive Workload</li> <li>Real User Actions</li> <li>Piloting</li> <li>System Response Time</li> <li>To Combine Qualitative Evaluation</li> <li>Collaboration with Real Users during Final Evaluation</li> </ul>	<ul style="list-style-type: none"> <li>Presence Questionnaire Scores</li> <li>Accuracy of Recommendations</li> <li>Precision</li> <li>Recall</li> <li>Accuracy of Retrieval</li> <li>Reliability Metrics</li> <li>Task Completion Time</li> <li>Task Effectiveness</li> <li>Task Efficiency</li> <li>Perceived Appropriateness of Adaptations</li> <li>Invasiveness of Adaptations</li> <li>Awareness of Adaptations</li> <li>Behavioural Complexity</li> <li>Accuracy of Models</li> <li>piA: Performance Influence on Adaptivity</li> <li>ApOC: Adaptive Personalisation Overall Cost</li> <li>Students' Content Learning Gains</li> <li>Presence Questionnaire Scores</li> <li>Intrinsic Motivation Inventory Scores</li> </ul>

Additional Notes or Comments

***D2.1.2 Provision of Explanations by EFEx Evaluation Framework***

Provision of Explanations on the Recommended Evaluation Techniques (Methods, Criteria and Metrics):

**Q12.** In this hybrid recommender system, we want to provide explanations on how the recommended techniques are derived. Do you think this would be a useful feature?

	Very Useful	Useful	Somehow Useful	Not Useful
Explanations on Recommended Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations on Recommended Evaluation Method (s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations on Recommended Measurement Criteria (s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations on Recommended Measurement Metrics(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations on Recommended Bundle (Method, Criteria and Metric)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

### **D3. Interacting with an Automated Hybrid (Case-based and Knowledge-based) Recommender System**

#### **D3.1 Novice Evaluators: Task-based Evaluation (Recommending Evaluation Techniques)**

Interacting with an Automated Hybrid (Case-based and Knowledge-based) Recommender System:

In this task you will be interacting with an automated hybrid (case-based and knowledge-based) recommender system. This system will recommend to you the most appropriate evaluation approach and techniques (methods, metrics and criteria) for evaluating an

adaptive system. The system will also a combination of the most appropriate (method + measurement criteria and metrics) which can be used together during evaluation of such a system. Throughout the recommendation process, you will be provided with explanations as to how the recommended techniques were derived.

Interacting with the Recommender System. After Interacting (i.e. using) the Recommender System. Please answer Questions 1 to 7

	Strongly disagree	Strongly agree
Q1. By using this system, can you more effectively identify the appropriate evaluation methods to be used when evaluating an adaptive system?		
Q2. By using this system, can you more effectively identify the appropriate measurement criteria to be used when evaluating an adaptive system?		
Q3. By using this system, can you more effectively identify the appropriate evaluation metrics to be used when evaluating an adaptive system?		

Recommender Effectiveness:

Q4. What features (i.e. characteristics) of the recommended evaluation techniques did you like most about the recommender system?

	Very Useful	Useful	Somehow Useful	Not Useful
Recommended Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Recommended Evaluation Method (s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Recommended Criteria (s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Recommended Evaluation Criteria(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A Combination of Recommended Techniques (Method + Criteria + Metric)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Q5. What features (i.e. characteristics) of the explanations did you like most about the recommender system?

	Very Useful	Useful	Somehow Useful	Not Useful
Explanations on Recommended Evaluation Approach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations on Recommended Evaluation Method (s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations on Recommended Measurement Criteria (s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations on Recommended Evaluation Criteria(s)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Explanations of Combination of Recommended Techniques(Method + Criteria + Metric)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Q6. Any Additional Comments (about the recommender system)

### D3.2 Novice Evaluators: Post Tasks Questions (SUS Questionnaire)

General User Satisfaction, Reaction and Comments (i...e After finishing interacting with the recommender System and completing Q1 to Q8 above)	
1. I think that I would like to use this recommender system frequently	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
2. I found the recommender system unnecessarily complex	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
3. I thought the recommender system was easy to use	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
4. I think that I would need the support of a technical person to be able to use this recommender system	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
5. I found the various functions in this recommender system were well integrated	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
6. I thought there was too much inconsistency in this recommender system	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
7. I would imagine that most people would learn to use this recommender system very quickly	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
8. I found the recommender system very cumbersome to use	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
9. I felt very confident using the recommender system	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree
10. I needed to learn a lot of things before I could get going with this recommender system	Strongly disagree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly agree

### D3.3 Results: A Summary of SUS Scores by 31 Novice Evaluators

Participant	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	SUS Score	Mean	SD
p1	5	2	5	1	5	1	4	2	4	1	90.0	3	1.67
p2	4	2	4	1	5	1	4	2	5	2	85.0	3	1.48
p3	5	1	4	2	5	1	3	2	4	1	85.0	2.8	1.54
p4	5	2	4	2	5	1	4	1	5	1	90.0	3	1.67
p5	5	1	4	2	5	1	4	1	4	1	90.0	2.8	1.66
p6	5	1	4	1	5	1	4	2	4	1	90.0	2.8	1.66
p7	5	1	4	1	5	2	4	2	4	2	85.0	3	1.48
p8	5	1	5	1	5	1	4	1	5	1	97.5	2.9	1.92
p9	5	1	4	2	5	1	4	2	4	1	87.5	2.9	1.58
p10	5	2	4	1	5	1	4	2	4	1	87.5	2.9	1.58
p11	4	1	4	1	5	1	3	2	4	1	85.0	2.6	1.50
p12	5	2	4	2	5	2	4	2	4	1	82.5	3.1	1.38
p13	5	1	4	2	5	1	4	1	4	1	90.0	2.8	1.66
p14	4	1	3	2	5	1	3	1	3	1	80.0	2.4	1.36
p15	3	2	3	3	5	1	3	2	3	2	67.5	2.7	1.01
p16	5	1	4	2	5	1	4	2	4	1	87.5	2.9	1.58
p17	5	1	5	1	5	1	4	1	4	1	95.0	2.8	1.83
p18	5	2	4	2	5	1	3	2	3	1	80.0	2.8	1.40
p19	4	2	3	2	5	1	3	2	3	2	72.5	2.7	1.10

Participant	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	SUS Score	Mean	SD
p20	4	1	4	1	5	2	4	2	4	1	85.0	2.8	1.47
p21	3	3	3	3	5	2	3	2	3	2	62.5	2.9	0.83
p22	5	1	4	5	4	1	4	2	4	2	75.0	3.2	1.47
p23	5	1	4	1	5	1	4	2	4	1	90.0	2.8	1.66
p24	4	2	4	2	4	2	4	2	4	1	77.5	2.9	1.14
p25	5	2	4	2	4	1	3	2	3	1	77.5	2.7	1.27
p26	4	2	4	2	5	2	4	2	4	2	77.5	3.1	1.14
p27	3	3	3	3	4	2	3	2	3	2	60.0	2.8	0.60
p28	5	1	5	1	5	1	5	1	4	1	97.5	2.9	1.92
p29	5	2	4	2	4	1	4	2	4	1	82.5	2.9	1.38
p30	2	3	2	4	3	1	2	1	2	3	47.5	2.3	0.90
p31	4	1	4	1	4	2	4	1	4	2	82.5	2.7	1.35

## D4. Experiment 2 - A Personalised Sub- Search System

### D4.1 Pre - Questionnaire

Interacting with A Personalised Search System that allows users to find evaluation studies  
 In this task you will be interacting with a personalised search system which searches across a knowledge repository of studies detailing the evaluation of adaptive systems, which were published between 2000 and 2012.

The system has three main features that allow novice and expert evaluators to search for:

- i) Find Evaluation Studies of Internal models of Adaptive Systems

ii) Find Evaluation Studies of Adaptive Systems.

iii) Find General Evaluation Studies of Adaptive Systems.

Q1. How familiar are you with personalised search systems (PIS) that allow to you find evaluation studies of adaptive systems?


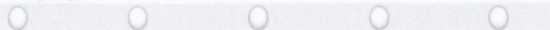

- a) Very Familiar       b) Familiar       c) Not Familiar

Q3. How often do you use personalised search system to find studies which detail evaluations of adaptive systems?


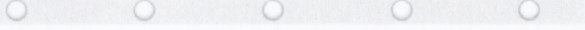

- a) Very Often       b) Regularly       c) Sometimes       d) Once or Twice       e) Never

Questions 4 to Q6 involves tackling (Task1, Task2 and Task3)

**Task1: Finding Evaluation Studies of Internal models of Adaptive Systems**

Q1. I found the Personalised Search System returned relevant search results for my query	<p style="text-align: center;">Strongly agree <span style="float: right;">Strongly disagree</span></p> 
Q2. I found the Personalised Search System returned irrelevant search results for my query	<p style="text-align: center;">Strongly agree <span style="float: right;">Strongly disagree</span></p> 
Q3. I found the presentation of the search results helpful	<p style="text-align: center;">Strongly agree <span style="float: right;">Strongly disagree</span></p> 

**Task2: Finding Evaluation Studies of Adaptive Systems (2000 to 2012)**

Q1. I found the Personalised Search System returned relevant search results for my query	<p style="text-align: center;">Strongly agree <span style="float: right;">Strongly disagree</span></p> 
Q2. I found the Personalised Search System returned irrelevant search results for my query	<p style="text-align: center;">Strongly agree <span style="float: right;">Strongly disagree</span></p> 
Q3. I found the presentation of the search results helpful	<p style="text-align: center;">Strongly agree <span style="float: right;">Strongly disagree</span></p> 



**Task3: Finding General Evaluation Studies of Adaptive Systems**

Q1. I found the Personalised Search System returned relevant search results for my query	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q2. I found the Personalised Search System returned irrelevant search results for my query	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q3. I found the presentation of the search results helpful	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree

Q3. What did you like most about the Personalised Search System?

- 1)
- 2)
- 3)
- 4)

Q4. What did you like least about the Personalised Search System?

- 1)
- 2)
- 3)
- 4)

Q5. Any additional comments?

## D4.2 Post -Tasks Questions (SUS Questionnaire)

General User Satisfaction, Reaction and Comments ( After finishing interacting with the recommender System and completing)

Q1. I think that I would like to use this Personalised Search System frequently	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q2. I found the recommender system unnecessarily complex	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q3. I thought the Personalised Search System was easy to use	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q4. I think that I would need the support of a technical person to be able to use this Personalised Search System	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q5. I found the various functions in this Personalised Search System were well integrated	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q6. I thought there was too much inconsistency in this Personalised Search System	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q7. I would imagine that most people would learn to use this Personalised Search System very quickly	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q8. I found the Personalised Search System very cumbersome to use	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q9. I felt very confident using the Personalised Search System	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree
Q10. I needed to learn a lot of things before I could get going with this Personalised Search System	Strongly agree <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> <input type="radio"/> Strongly disagree

### D4.3 Results: A Summary of the SUS Scores by the 33 Participants

Participant	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	SUS Score	Mean	SD
p1	4	2	4	2	3	2	3	3	4	2	67.5	2.9	0.83
p2	3	2	3	2	4	1	3	2	4	1	71.5	2.5	1.03
p3	5	2	4	2	4	1	5	5	5	1	80	3.4	1.63
p4	4	2	5	1	5	2	4	2	3	3	77.5	3.1	1.30
p5	5	2	5	1	4	3	4	1	5	2	85	3.2	1.54
p6	2	2	3	1	5	1	5	1	5	1	85	2.6	1.69
p7	4	1	5	1	5	1	5	1	5	2	95	3.0	1.84
p8	4	2	4	1	4	1	4	3	5	2	80	3.0	1.34
p9	5	4	3	4	3	3	4	2	4	3	57.5	3.5	0.81
p10	3	3	3	2	4	5	5	4	4	2	57.5	3.5	1.03
p11	3	3	3	2	3	5	4	4	4	2	52.5	3.3	0.90
p12	4	2	5	1	4	1	5	3	4	1	85	3.0	1.55
p13	3	1	4	1	3	2	3	3	4	3	67.5	2.7	1.01
p14	3	2	4	1	3	1	4	2	3	2	72.5	2.5	1.03
p15	4	1	4	1	3	1	5	1	4	3	82.5	2.7	1.49
p16	2	4	3	4	2	2	3	4	3	4	37.5	3.1	0.83
p17	4	2	4	1	4	2	2	3	4	4	65	3.0	1.10
p18	3	3	2	4	3	4	1	4	4	1	42.5	2.9	1.14
p19	4	3	3	2	3	2	4	4	4	5	55	3.4	0.92
p20	4	3	4	4	4	2	3	2	4	4	60	3.4	0.80
p21	3	3	4	1	4	2	5	2	5	1	80	3.0	1.41
p22	3	4	4	2	4	3	4	2	3	4	57.5	3.3	0.78
p23	4	1	5	2	4	2	4	2	5	1	85	3.0	1.48
p24	4	1	4	1	4	3	2	2	4	1	75	2.6	1.28
p25	4	2	4	1	5	1	5	1	4	1	90	2.8	1.66
p26	4	3	3	2	4	3	5	3	3	2	65	3.2	0.87
p27	3	2	5	2	4	2	4	2	5	1	80	3.0	1.34
p28	4	3	3	4	4	3	5	3	3	3	57.5	3.5	0.67
p29	3	5	2	4	3	3	4	3	2	4	37.5	3.3	0.90
p30	4	2	4	2	3	2	3	2	4	2	70	2.8	0.87
p31	4	3	4	4	3	2	3	3	4	3	57.5	3.3	0.64
p32	3	3	3	2	4	2	3	2	2	4	55	2.8	0.75
p33	3	1	4	3	3	3	4	4	4	4	57.5	3.3	0.90

## D5. Experiment 3 - A Taxonomy of Technical Terms

### D5.1 Identification of User Characteristics

C1. How familiar are you with Taxonomies of Evaluation approaches used and Techniques (methods, criteria and metrics) for Adaptive Systems

- a) Very Familiar       b) Familiar       c) Not Familiar

Q2. How often do you use Taxonomies of Evaluation approaches used and Techniques (methods, criteria and metrics) for adaptive systems that helps non expert evaluators to understand different aspects of the evaluations of such systems

- a) Very Often       b) Regularly       c) Sometimes       d) Once or Twice       e) Never

A Taxonomy of Evaluation approaches used and Techniques (methods, criteria and metrics) for Adaptive:

Q3. What did you like most about the taxonomy?

Q4. What did you like least about the taxonomy?

Q5. Any additional comments?

## D5.2 Post -Tasks Questions (SUS Questionnaire)

General User Satisfaction, Reaction and Comments (i.e. After finishing interacting with the taxonomy of technical terms)

<p><b>Q1.</b> I think that I would like to use this taxonomy</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q2.</b> I found the taxonomy unnecessarily complex</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q3.</b> I thought the taxonomy was easy to understand</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q4.</b> I think that I would need the support of a technical person to be able to use this taxonomy</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q5.</b> I found the various functions in this taxonomy were well integrated</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q6.</b> I thought there was too much inconsistency in this taxonomy</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q7.</b> I would imagine that most people would learn to use this taxonomy very quickly</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q8.</b> I found the taxonomy very cumbersome to use</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q9.</b> I felt very confident using the taxonomy</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 
<p><b>Q10.</b> I needed to learn a lot of things before I could get going with this taxonomy</p>	<p>Strongly agree <span style="float: right;">Strongly disagree</span></p> 

### D5.3 Results: A Summary of the SUS Scores by the 15 participants

Participant	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	SUS Score	Mean	SD
p1	5	2	5	1	5	1	5	1	5	1	97.5	3.1	1.92
p2	4	3	4	2	4	2	4	2	4	2	72.5	3.1	0.94
p3	5	2	5	1	5	1	5	1	5	1	97.5	3.1	1.92
p4	3	1	3	1	3	2	3	2	3	4	62.5	2.5	0.92
p5	5	2	5	1	5	1	5	1	5	1	97.5	3.1	1.92
p6	4	1	4	1	4	1	4	1	4	1	87.5	2.5	1.50
p7	4	1	4	2	4	1	4	1	4	1	85	2.6	1.43
p8	5	2	5	1	5	1	5	1	5	1	97.5	3.1	1.92
p9	4	1	5	1	5	1	5	1	5	1	97.5	2.9	1.92
p10	3	1	3	2	3		3	1	3	5	60	2.6	1.11
p11	4	1	4	1	4	1	4	1	4	1	87.5	2.5	1.50
p12	4	2	5	1	5	1	5	1	5	1	95	3	1.84
p13	4	2	5	1	5	1	5	1	5	1	95	3	1.84
p14	3	2	4	1	3	1	3	1	3	1	75	2.2	1.02
p15	4	1	4	1	4	1	4	1	4	1	87.5	2.5	1.50