Towards comparative evaluations of user-adaptive software systems

Athanasios Staikopoulos and Owen Conlan
The ADAPT Centre
School of Computer Science and Statistics
Trinity College Dublin
{Athanasios.Staikopoulos, Owen.Conlan}@scss.tcd.ie

ABSTRACT
Currently there are many user-adaptive systems across different application types including education, business and information systems. However, evaluating user-adaptive systems is still a challenging research issue and a difficult task. This is because of the lack of widely accepted evaluation methods, data and the difficulty on generalizing the application areas. In order to provide comparative and convincing arguments that the adaptation works and it is valuable to the users, it is vital to ensure a scientific process and to provide common evaluation mechanisms, methods and metrics. Taking inspiration from similar research communities and building upon current evaluation approaches, a question arises as to whether it is possible to create such a community and develop a common evaluation mechanism of user-adaptive systems. In this paper, we indicate how we can move towards comparative evaluations of user-adaptive systems.

Keywords
User-adaptive systems; personalization; evaluation

1. INTRODUCTION
User adaptive software systems [1] are interactive systems that modify their properties and behaviors to the different requirements and changing needs of individuals or groups of users, over time and context. Currently there are many user-adaptive systems for different application types such as for education, business, information systems, health and digital humanities. More likely such adaptive systems model users based on different information, perform adaptations across different aspects (dimensions) and use techniques and algorithms from different domains. For example Information Retrieval (IR), Machine Learning, Data Mining, Artificial Intelligence, Human Computer Interaction, Semantic Web and Rule based systems.

Even so, some general functionalities and features are expected to be present in any user-adaptive system upon which common evaluations can be designed and performed. Having a common evaluation mechanism that could be used to evaluate and compare different user adaptive systems and approaches both comprehensively and upon specialized aspects, is very beneficial on developing optimized and effective adaptation solutions. In general the evaluation and comparison of user-adaptive systems is considered an important and challenging research issue. This is because of the lack of widely accepted evaluation methods, data and the difficulty on generalizing the application types. Taking inspiration from similar research communities on Information Retrieval and Machine Translation, user-adaptive systems could benefit from such experience and try to develop similar communities and evaluation mechanisms.

2. TOWARDS COMPARATIVE EVALUATIONS
In order to produce effective comparative evaluations of user-adaptive systems first there is a need to specify an evaluation methodology with detailed guidelines, evaluation models, metrics, and standardized questionnaires. Next, similar to other research communities we need to establish a related community that will design common shared tasks and scenarios for general and application specific types as well as different adaptation aspects. In addition, test content collections will need to be provided in order to evaluate adaptations with controlled experiments. Finally, access to the evaluation results, data sets, and tools will need to be provided via the community portal. As a result the evaluation approaches will be evaluated and compared from the same communities, on the same tasks, using the same evaluation models and metrics. In addition, both automated methods and human based evaluation methods (quantitative and qualitative) will be needed.

Current evaluation approaches to adaptive systems include **User-Centered evaluations** [2] that are based on verifying the observable/perceived qualities of the adapted system, in particular related to the user satisfaction. **Empirical evaluations** [3] that refer to the appraisal of a theory by observation in experiments. Such evaluations are based on the design and execution of controlled experiments so that the particular factors that are to be tested are separated from other confounding factors. They are particular helpful to determine and evaluate user adapted interactions. Finally, the **Layered evaluation** [4][5] approaches that do not treat evaluations as a monolithic process, instead the main phases of the adaptation process are divided into separate layers which are then evaluated individually. This approach has proven useful to identify and evaluate the exact characteristics of a user-adaptive system.

2.1 Decomposing Adaptation
Figure 1 below depicts how the adaptation process is separated (decomposed) into several layers, so a holistic evaluation approach can be performed across the different adaptation phases. Specific evaluation properties and metrics are captured and effectively evaluated for each layer. As a result, different adaptive systems can be compared with each other upon a specific layer and upon common evaluation criteria. Adaptive systems do not have to provide implementations for every layer, feature or for every metric. In general, user-adaptive software systems can be decomposed and evaluated upon the following dimensions (layers):

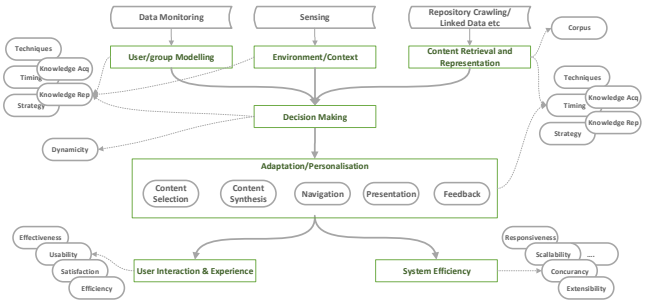

Figure 1. Decomposing the adaptation process

2.1.1 User and Group Modelling

In general the User and Group Modelling dimension captures the preferences, interests, needs, prior knowledge, roles and goals of users. Specific application types can capture additional data, for example in educational systems we may want to capture the user's previous modules, grades and attendance. Similarly, we can capture details about the formed communities (groups). The model acquisition can be manual (e.g. pre-configured), automated (extracted and build automatically) or semi-automated. A number of different methods can be used to classify users, identify the user's intend and build communities with machine learning, data mining and recommender techniques.

2.1.2 Environment and Context

This dimension considers a number of external environmental factors that can affect the adaptation such as location, time, language, and device. It also refers to situation-adaptive systems that can sense or predict the human affect or mental states in varying contexts and accordingly adapt the system. For example predicting that the user is tired, bored, confused, if he is at work, travelling, or at home. Different techniques are applied such as Bayesian networks (BNs) - graphical models that represent the probabilistic relationships among variables of interest.

2.1.3 Content Retrieval and Representation

This dimension refers to the evaluation of Content retrieval and its potential representation with a domain model, topic map, etc. Typically approaches are from the field of Personalized Information Retrieval [6], Natural Language Processing, Recommender systems, Query expansion, Linked data, semantic and ontology matchmaking. Content can be retrieved either from a closed or an open corpus.

2.1.4 Adaptation Decision Making

Even if the system has inferred the user properties, there is a need to evaluate the actual decision making mechanism (e.g., adaptation strategy, adaptation model) and check if the most optimal adaptation was selected based on the current criteria. Decision making can be based on different techniques from rule based systems, to decision trees and Neural Networks. In general the adaptation decision making can be characterized as static, dynamic or autonomous. Static are the ones that are specified and fixed at design time. Dynamic adaptations are the ones that are performed on the fly (run time) due to changes on the user models, requirements. Autonomic, are the ones that automatically adjust and decide the best way to accomplish a given strategy despite the changing environmental conditions and demands.

2.1.5 Adaptation

This dimension evaluates the adaptation effect after applying a specific method/technique. The following major categories are considered: **Content Selection & Recommendation** refers to the actual selection of content and recommendations. A number of different techniques can be applied from IR, query expansion, scoring, recommenders, semantic matchmaking etc. The evaluation metrics are related to the ones used in IR evaluations such as precision, recall, f-measure, Mean Average Precision, Discounted Cumulative Gain etc. **Content Composition** refers to the composition and generation of content from fragments. **Navigation** refers to the sequencing of content, topics or links. Different techniques are used to generate sequences (e.g. rule based, constraint satisfaction, planning) as well as to hide, remove, disable and annotate links (e.g. human computer interaction). **Presentation** refers to how we present (style and layout) certain components to users based on their profile, environment and context (handheld device). **Support and Feedback** [7] refers to the supplementary mechanisms used to enhance the user's experience for example by providing feedback, guidance, notifications, reminders etc.

2.1.6 User Interaction and Experience

In this case the users evaluate the user-adaptive system in terms of the perceived user experience and interaction. For example, the **usability** of the system is often evaluated by user-centered approaches and metrics. **Satisfaction** is also examined by direct user feedback. **Effectiveness** evaluates the effectiveness of adaptivity performed by the system. Different metrics are used based on the nature and use of the system for example knowledge gain and retention for educational systems, or needed information for information systems [8]. **Efficiency** is also often used to measure the time or the number of steps taken to complete a task.

2.1.7 System Properties and Efficiency

This dimension contains features that address the system from a software-engineering perspective. In particular, we should consider **Responsiveness** referring to the time taken the adaptive system to implement a specific adaptation. **Scalability** that evaluates the degree to which the system could be scaled both in terms of the users and the content size. **Concurrency** that evaluates the degree to which the system can accommodate the requests of concurrent users. **Extensibility** to if the system can be easily extended across different dimensions and features. **Openness** if the system provides access to its code, algorithms and services, so it could be extended and evaluated independently. Development **Cost** to model and realize a particular adaptation shared task or scenario. Ideally this can be evaluated and measured during a community workshop session. Finally, **security** to ensure that the system provides secured access and **privacy** so the system protects user's privacy, identity and data.

2.2 Evaluation Criteria/Metrics

The adaptation dimensions and aspects can be related to the following evaluation criteria and metrics.

- Corpus (open, closed)
- Knowledge Acquisition (manual, automated, semi-automated)
- Knowledge representation (xml, ontology, text)
- Dynamism (static, dynamic, autonomic)
- Strategy (reactive, proactive)
- Timing (event, scheduled, interval)
- Techniques (Machine Learning, NLP, Ubiquitous Computing, IR, AI, Rule Based Systems, Complex Systems, User interface, Mining, Recommenders, etc)

3. CONCLUSIONS

Evaluating adaptive systems is a difficult process. In this paper, we have indicated the core requirements in order to provide

effective and comparative evaluations of user-adaptive systems. More specifically we have proposed developing a flexible common evaluation model that could be used to evaluate and compare different user-adaptive systems and approaches both comprehensively (as a whole) as well as upon specific adaptation layers and aspects. Finally, we emphasized the need of establishing a related research community that would design common shared tasks and scenarios for evaluating the adaptive systems with a combination of evaluation methods such as user-centered, empirical, layered and heuristic.

4. ACKNOWLEDGMENTS

This research is supported by the Science Foundation Ireland (12/CE/I2267) as part of the ADAPT Centre (www.adaptcentre.ie) in Trinity College Dublin.

5. REFERENCES

[1] M. Schneider-Hufschmidt, U. Malinowski, and T. Kuhme, *Adaptive User Interfaces: Principles and Practice*. 1993.

[2] L. Van Velsen, T. Van Der Geest, R. Klaassen, and M. Steehouder, "User-centered evaluation of adaptive and adaptable systems: a literature review," *Knowl. Eng. Rev.*, vol. 23, no. 03, pp. 261–281, 2008.

[3] D. N. Chin, "Empirical Evaluation of User Models and User-Adapted Systems," *User Model. User-adapt. Interact.*, vol. 11, no. 1–2, pp. 181–194, 2001.

[4] P. Brusilovsky, C. Karagiannidis, and D. Sampson, "The Benefits of Layered Evaluation of Adaptive Applications and Services," *World Wide Web Internet Web Inf. Syst.*, pp. 1–8, 2001.

[5] S. Weibelzahl and G. Weber, "Advantages, opportunities and limits of empirical evaluations: Evaluating adaptive systems," *Ki*, vol. 16, pp. 17–20, 2002.

[6] B. Steichen, H. Ashman, and V. Wade, "A comparative survey of Personalised Information Retrieval and Adaptive Hypermedia techniques," *Inf. Process. Manag.*, vol. 48, no. 4, pp. 698–724, 2012.

[7] A. Staikopoulos, I. OKeeffe, B. Yousuf, O. Conlan, E. Walsh, and V. Wade, "Enhancing Student Engagement through Personalized Motivations," *2015 IEEE 15th Int. Conf. Adv. Learn. Technol.*, pp. 340–344, 2015.

[8] S. Lawless, A. O. Connor, and C. Mulwa, "A Proposal for the Evaluation of Adaptive Personalized Information Retrieval," pp. 1–4, 2010.