Conversational Evaluation of Personalized Solutions for Adaptive Educational Systems

Martin Labaj, Mária Bieliková

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
{labaj,bielik}@fiit.stuba.sk

Abstract. Current educational systems offer many personalized features. But do they really help? Which personalization method is the best in given settings? Evaluating the personalization in an educational system is as important as designing the methods themselves. While many quantitative and qualitative methods have been explored previously, there are various rules and issues when performing experiments with participants – users. E.g. users should not interact with experimenters, but on the other hand, only post-session testing can be affected by maturation. We proposed a conversational evaluation approach in which we combine advantages of uncontrolled experiments with advantages of other methods. We describe a method for evaluation questions asked to the users in appropriate moments during their work in a web-based environment and its initial evaluation within ALEF adaptive learning framework.

1 Introduction and Related Work

When designing and operating adaptive and personalized systems in particular, evaluation is an essential part of the process. Without knowing the performance of the personalized system when employing various methods with various parameters, it is impossible to judge its effectiveness, pick successful methods, or make adjustments at all. Differing evaluation methods can be employed for one adaptive system, even on multiple levels, evaluating the adaptive system by parts [1], where we can often omit users, e.g. by using golden standards. Here we focus on those evaluation methods, which include users using the adaptive system being evaluated.

One can focus on if and how users interact with the experimenter in a user-centered evaluation and recognize several common approaches [1, 2]: *questionnaires* (series of questions displayed on paper or within a system), *interviews* (interviewer asks the user), *data log analysis* (user actions are recorded and analyzed without the participation of the user, *focus groups* (a discussion in a group of participants), and *think-aloud protocols* (the user describes their actions during the session).

Regardless of whether the users are interviewed, given pre-tests or questionnaires, or we only use logs of their activity, another categorization can made on how the experiment is performed. Three types of experiments are a common practice in adap-

tive systems such as recommender systems [3]: *offline* evaluation, *user studies* (*controlled experiments*), and *online* evaluation (*uncontrolled experiments*).

In *offline* experiments, previous user interaction with the system is recorded and used without the users. For example, we can record which learning objects were visited by users and how they were rated, predict user ratings using collaborative filtering and evaluate from the recorded data whether the user actually rated the learning objects as predicted. In a *user study*, a group of users in a controlled environment work with the system. The user feedback can be gathered using multiple methods – e.g. with think-aloud protocols during the session, while also using post-session questionnaires. In *online* experiments (where "online" does not indicate network environment, but rather live system being used), users work towards their goals in their own settings, for example a recommender system is deployed into live learning system.

There are rules which should be observed whenever possible, ranging from random assignment of participants, through instructing them in the same way, to maintaining uniform work environment [4]. These rules are aimed at leveling out the effect of nuisance variables and can be obeyed either by accordingly preparing the test room, written instructions, etc., or also by randomizing their influence, e.g. letting large volume of users work in their own environment at own times (naturally, doing the same for experiment and control groups).

One decision is that we either let the users go alone without any influences, and in fact, large number of users can participate this way, obtaining vast datasets for numerical evaluations, or we bring the users into laboratory, where we can have absolute overview of their actions, expressions, etc. and even talk to them (think-aloud methods). Whether we have the user at hand in laboratory, or perform the experiment remotely, when we seek out opinions from the users, or, in a learning system, want to assess their knowledge for evaluation using the measure of gained knowledge, we have several options. We can passively provide commenting or rating tools in the system and count on the users using them. We can interact with the user during the work (e.g. think-aloud), but this can alter the user's behavior, e.g. the user can approach problems differently when speaking out loud about them [4]. Or we can use post-testing, but we will maybe introduce maturation factors (users forget and also after the whole session, they can look differently on specific events).

We seek for a method of user-centered adaptive educational systems evaluation combining advantages of the above – capturing the user feedback right during users' workflow, but with minimal impact on learning process, i.e. without interrupting them significantly, and doing so in their natural settings. We proposed a conversational evaluation approach using *evaluation questions* being displayed at the appropriate times for collecting explicit user feedback. In this paper, we describe the evaluation questions approach and its use within our Adaptive LEarning Framework ALEF.

2 Conversational Evaluation

In our conversational evaluation approach, we generate *evaluation questions* and display them in appropriate moments during user's normal unsupervised work. Using

evaluation questions is not unfamiliar to user feedback elicitation approaches, such as recommender rating elicitation [5], where an adaptive system asks the user for their ratings as needed, e.g., when the recommender does not have enough information to pick an item to be recommended for the given user.

Asking for the item rating instead of waiting for the user to provide one has several advantages: the users can be more motivated to even provide the rating in the first place, as they are told that they are helping improving their profile and helping the system with recommendation to others [5], the user can achieve higher satisfaction and perceive the system as more useful [6], and of course accuracy can be increased while decreasing the load on user, as ratings are elicited for such items whose the rating or re-rating would be useful to the system [7].

We follow similar line with conversational evaluation questions – instead of waiting for the user to provide the feedback after the session through the post-test/questionnaire, or on the other side, asking them to unnecessarily comment every thought, relevant evaluation questions are asked at appropriate moments. A user of an adaptive educational system can be sitting at home, studying for the next exam and at the same time helping evaluating/improving the system and associated personalization methods by answering sparsely displayed questions.

We propose a rule-based framework for generating conversations aimed at evaluation of user opinion, knowledge, etc. Evaluation questions are based on classic question types: yes/no, single choice, multiple choice, or free text. The text of questions is prepared by educational system developers (designer of particular personalization technique being tested) and stored as *questions templates*. When a question template is selected by the question engine and adapted to the user and situation (by processing the template scripts), it becomes a *question instance*, asked to a given user, within a given setup (e.g. learning course), and comes from a given asking rule.

The evaluation questions (their respective templates) are selected by question asking engine based on *triggers*. The triggers are composed of a rule-part with arbitrary conditions to be evaluated against the user model and of pre-assigned question template. For example, when a user scans through a list of items in a navigation tool (providing recommendations) and then proceeds to use the non-adaptive menu instead, a question template on why the tool was not used is triggered. The triggers have pre-assigned priorities and the most immediate question template takes over. As we are aiming at web-based systems, the triggers can be based on client-side user actions, as well as server side logs. Fig. 1 shows an overview of the involved entities.

The questions can be asked in a *synchronous* elicitation, which occurs during or just after an action, e.g. asking the user to rate a learning object after they finished reading this object, and in an *asynchronous* elicitation, which can, for example, occur when a user input is needed regardless of his actions.

The selected and instantiated question is displayed to the user in the foreground, darkening the rest of the system screen. In order not to obtain random answers (just for the window to go away), the user can chose not to answer this question, or to answer it later (if feasible due to the nature of the question). All these steps – from selecting the question template using triggers, to instantiating the question template, to asking and possibly re-asking the question, to receiving the answer(s), are logged.

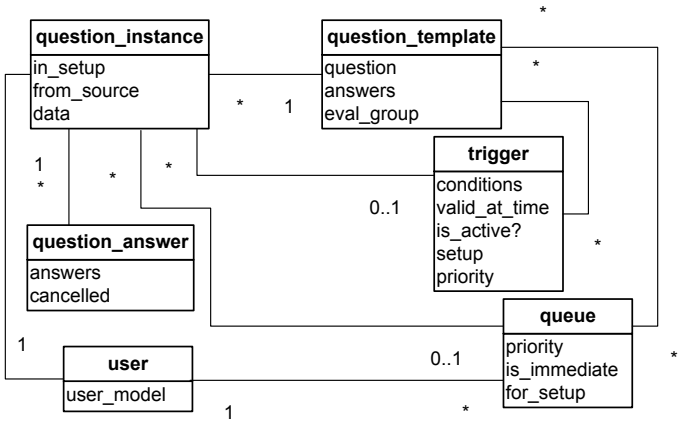

Fig. 1. Conceptual overview of the evaluation question model.

3 Evaluation Questions within ALEF Learning Framework

We realized the proposed evaluation questions approach as an evaluation sub-framework within the ALEF (Adaptive LEarning Framework) [8]. It is being used on the Faculty of Informatics and Information Technologies for five years in several courses: Principles of Software Engineering, Functional Programming and Logic Programming, and Procedural Programming. ALEF (Fig. 2) offers the users (students) learning objects of Explanation, Exercise, Question, and Question-Answer type, and also a variety of tools, ranging from personalized recommendations of learning objects, to automatic content enrichment using external sources. ALEF also provides domain and user models, as well as other features necessary for such personalized solutions. The educational system is used by students both during lectures in supervised settings in laboratories, as well as at home, unsupervised.

One example of a feature problematic to evaluate is the recommender system usage. When the user follows a recommendation to study the proposed learning object as next, it can be evaluated through measures such as time spent (immediate return is a negative indicator, staying for some time is a positive indicator that the user has liked the recommendation), through subsequent user rating of the item, or through adaptive knowledge testing via exercises and questions (user's knowledge has stayed the same or increased). However, when the user does not follow any recommendation (and this is a frequent case), it can have various meanings – maybe the user did not notice the recommendations, or he does not understand them, or the recommendations are inaccurate so the user ignores them, or the user just do not want to use them at all for own reasons often related to personal goals and motivational elements for using the educational system. Do we need to make the recommendations visually more prominent in the system? Or explain them in a better way? Employ different recommendation method? Questions like these can be answered by prompting the user non-invasively during his focus changes when using other tools in the system to navigate (e.g. ask: Why did he choose the menu over the recommended items?).

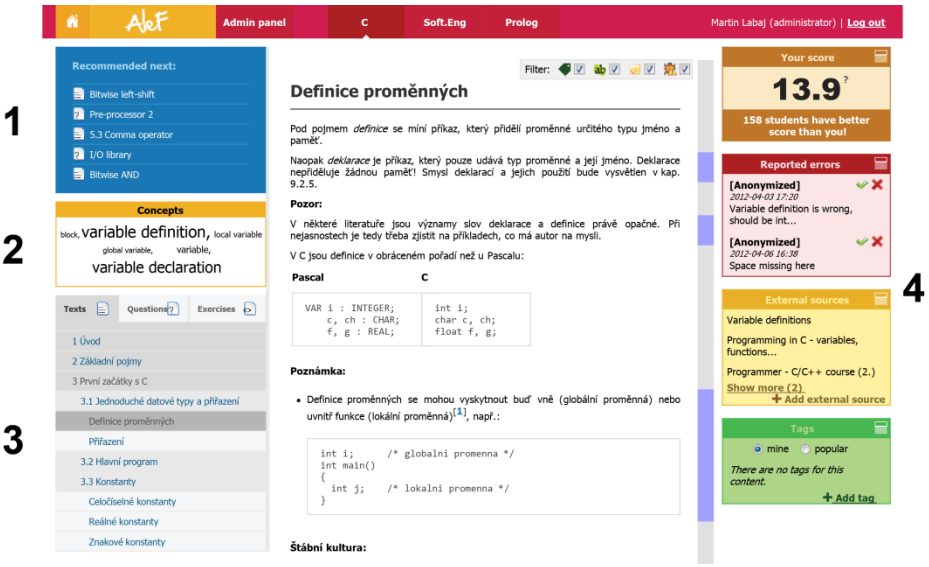

Fig. 2. Screenshot of ALEF educational system, showing a learning object ("Variable definition", in Slovak) in the middle. The left side contains tools for navigation between learning objects: personalized recommendations (1), tag recommendation (2), menu (3). The right side contains tools within the learning object (4): reported errors, external sources, tags.

4 Evaluation and Conclusions

Our rule-based evaluation question framework was used in several evaluations, most recent on summarization of explanation parts of learning objects [9]. Here we present a case of synchronous questions based on user attention. 34 students took part in two week uncontrolled experiment. User attention was tracked using mouse interaction and commodity gaze tracking via webcams and based on attention, application (tools) and document (learning object) fragments were assessed and recommended.

We hypothesized in this experiment that when asked at the appropriate moments, the users are more willing to provide opinions. The questions were displayed by the question engine in two situations. When the user focused on a fragment for a period of time and then shifted focus away, a question about the fragment was instantiated. The same question templates were also triggered randomly, unrelated to user attention, i.e. even during focused work, and unrelated to their current target fragment. In each question, the users had access to afore mentioned options of postponing the question or declining to answer. When asking questions related to user's previous fragment and in the moments of shifting the focus, only 7 % (using gaze and mouse) and 12 % (mouse only) of the instantiated questions were cancelled. Contrast to this, 33 % of randomly displayed questions was dismissed.

Our experiment suggests that when the evaluation questions are asked at the appropriate time and right when the user is working with the part in question (or just finished working with it), we can accomplish higher cooperation from the user

providing more feedback than when we would ask them randomly. This is not the only advantage. Since the users provide their opinions during their work, right when they interact with a given object, such as recommendation, they provide higher quality feedback than when commenting/rating after the entire session, and yet, they are not as interrupted as in supervised think-aloud evaluation.

Our approach does not aim to entirely replace the physical presence of the user and the possibility of observing directly what are they doing and asking additional, unprepared questions. It is rather a supplement, which gives different views on adaptive mechanisms and collects such views from users participating even in an unsupervised online experiment. Although the questions are constructed when asked and adapted to the user, some pre-thought is needed to create the templates in advance.

Currently we are interested in combining the conversational evaluation with the elicitation of ratings, especially based on user attention and also in continuous adaptive testing of user knowledge, using these mechanisms with questions created by the teacher or sourced from the learning content.

Acknowledgements: This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11 and by the Slovak Research and Development Agency under the contract No. APVV-0233-10.

References

1. Paramythis, A., Weibelzahl, S., Masthoff, J.: Layered evaluation of interactive adaptive systems: framework and formativemethods. User Modeling and User-Adapted Interaction. 20, 383–453 (2010).
2. Van Velsen, L., Van Der Geest, T., Klaassen, R., Steehouder, M.: User-centered evaluation of adaptive and adaptable systems: a literature review. The Knowledge Engineering Review. 23, 261–281 (2008).
3. Shani, G., Gunawardana, A.: Evaluating Recommendation Systems. In: Ricci, F., Rokach, L., Shapira, B., and Kantor, P.B. (eds.) Recommender Systems Handbook. pp. 257–297. Springer US, Boston, MA (2011).
4. Chin, D.: Empirical evaluation of user models and user-adapted systems. User modeling and user-adapted interaction. 181–194 (2001).
5. Carenini, G., Smith, J., Poole, D.: Towards more Conversational and Collaborative Recommender Systems. Proceedings of the 8th international conference on Intelligent user interfaces - IUI ’03. pp. 12–18. ACM Press, New York, New York, USA (2003).
6. Knijnenburg, B.P., Willemsen, M.C.: Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. Proc. of the third ACM conf. on Recomm. sys. - RecSys ’09. p. 381. ACM Press, New York, USA (2009).
7. Amatriain, X., Pujol, J.M., Tintarev, N., Oliver, N.: Rate it Again: Increasing Recommendation Accuracy by User re-Rating. Proc. of the third ACM conf. on Recommender. systems - RecSys ’09. pp. 173–180. ACM Press, New York, USA (2009).
8. Šimko, M., Barla, M., Bieliková, M.: ALEF: A framework for Adaptive Web-Based learning 2.0. Key Competencies in the Knowledge Society, WCC '10. pp. 367–378 (2010).
9. Móro, R., Bieliková, M.: Personalized Text Summarization Based on Important Terms Identification. 23rd Int. Workshop on Database and Expert Systems Applications. pp. 131–135. IEEE (2012).