User-Centric Evaluation Framework for Multimedia Recommender Systems

Bart Knijnenburg[1], Lydia Meesters[1], Paul Marrow[2], and Don Bouwhuis[1]

[1] Technische Universiteit Eindhoven
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
{B.P.Knijnenburg,L.M.J.Meesters,D.G.Bouwhuis}@tue.nl
[2] BT Innovate
Orion 3 Room 7 PP 1, Ipswich, IP5 3RE, UK
Paul.Marrow@bt.com

Abstract. Providing useful recommendations is an important challenge for user-centric media systems. Whereas current recommender systems research mainly focuses on predictive accuracy, we contend that a truly user-centric approach to media recommendations requires the inclusion of user experience measurement. For a good experience, predictive accuracy is not enough. What users like and dislike about our systems is also determined by usage context and individual user characteristics. We therefore propose a generic framework for evaluating the user experience using both subjective and objective measures of user experience. We envision the framework, which will be tested and validated in the large-scale field trials of the FP7 MyMedia project, to be a fundamental step beyond accuracy of algorithms, towards usability of recommender systems.

1 Introduction

The common practice in recommender systems research mainly focuses on algorithm performance [9]. Measures like mean absolute error or root mean square error measure the accuracy of an algorithm in predicting user ratings [3]. The implicit assumption is that better predictions lead to better recommendations, and that better recommendations lead to higher user satisfaction.

There are several flaws in this approach to recommender system evaluation. First of all, algorithm accuracy is not necessarily correlated with user experience [6]. Furthermore, the user experience of an interactive product such as a recommender system does not depend exclusively on prediction accuracy but is also influenced by factors such as the ease of use, usefulness, engagement, transparency, satisfaction, novelty and enjoyment of the system itself and the items retrieved [4, 10]. This list of factors is far from complete, and current literature lacks a unified understanding of what experience factors need to be measured or how to measure them. Although it is commonly understood that user experience is best measured using a combination of objective (user behaviour) and subjective (questionnaire) measures [2], there have been few attempts at a structural approach towards user experience measurement for media recommenders [7]. A framework is therefore needed for evaluating the user experience of media recommender systems. Specifically, the framework should:

P. Daras and O. Mayora (Eds.): UCMedia 2009, LNICST 40, pp. 366–369, 2010.

- provide a generic set of measurable concepts related to experience that is applicable to all types of media recommender systems,
- link the subjective experience measurements to objective recommender system aspects on the one hand, and objective user behaviour on the other hand,
- take into account personal and situational characteristics that may also influence the user experience [4, 8, 9],
- be validated in both qualitative studies (user diaries, interviews and focus groups) and quantitative studies (questionnaires).

2 Evaluation Framework

In order to construct our user-centric evaluation framework, we take the proven concept of the technology acceptance model (TAM) as a starting point [1], but also include hedonic experiences such as appeal, fun, and user emotions [2]. The basic framework, shown in figure 1, covers objective recommender system aspects, subjective evaluations, subjective experiences and objective behaviours, and relates them in a structural way. Furthermore, as the particular situation and the particular user may matter for the evaluation, it embeds the framework in situational and personal characteristics. A brief description of each component is given below, for a detailed description see [7].

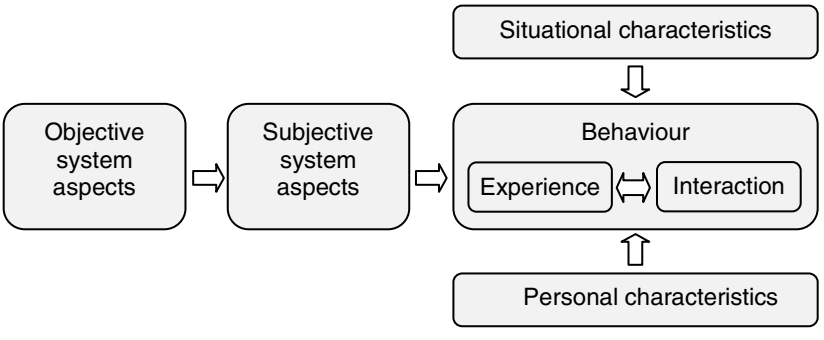

Fig. 1. User-centric evaluation framework for Multimedia recommender systems

The main assumption in our framework is that the *objective system aspects* (qualities of the recommender system) eventually influence the user experience [9]. These aspects include the visual and interaction design, the recommender algorithm, the presentation of recommended items, explanations and additional system features such as social networking and profile control. To measure the impact of a specific system aspect on the user, one would ideally manipulate the aspect across two or more conditions. For example, to test the impact of profile control, one would test a system that provides no profile control against a system that provides some profile control. By keeping all other system aspects the same, one can observe the effect of the manipulated aspect on the users' perception, experience and behaviour.

The *subjective system aspects* are users' evaluations of the objective system aspects. Although these evaluations may differ between users, they should typically not be influenced by personal characteristics and situations. Subjective system aspects include system usability, perceived item quality and visual attractiveness.

The *situational* and *personal characteristics* are beyond the influence of the system itself, but may significantly influence the user experience and interaction. In different situations users may use the system for different tasks, and in these tasks the choice goal and the domain knowledge may differ [4]. Personal characteristics that are typically stable across tasks include trust (in general, not in terms of the system), control and social factors [8, 9].

The *experience* also represents an evaluation of the system by the user. However, in contrast to the subjective system aspects, it is the users' behaviour and not the system that is the main focus of the experience measures. The experience typically depends on personal characteristics as the system aspects may influence different users in different ways. Also, the experience may change over time and across different choice situations. Experience is divided conceptually between hedonic qualities, usefulness, trust and outcome evaluations [2].

The *interaction* itself can also become an object of investigation of recommender systems research. Interaction is objectively measurable, and determines the 'final step' of the actual evaluation of the system: a system that is evaluated positively will be used more extensively. We explicitly link the objective interaction to objective system aspects through a series of subjective constructs, because these constructs are likely to attenuate and qualify the effect of the objective system aspects on the interaction [9]. Perception and experience explain why users' behaviour is different for different systems. This explanation is the main value of the user-centric evaluation of recommender systems [6].

3 Future Work

The evaluation framework is a conceptual model that still needs to be tested and validated. Currently, on-line tests in real-world settings are being deployed and the first results will be available in the next weeks. The initial aim is to systematically manipulate particular recommender system features and measure their effect on selected components in the framework. Specifically we are currently testing:

- the effect of personalized recommendations on the users' evaluation, experience and interaction with a media recommender system,
- the effect of recommendation variety on the users' evaluation, experience and interaction with a media recommender system,
- the effect of recommendation quality on the users' willingness to share more information with the system,
- the effect of personal trust factors on the willingness to share information with a media recommender system.

These effects are tested in on-line experiments across three different MyMedia recommender system prototypes in three different countries. The results will be used to refine the framework and to develop a set of evaluation metrics that can discriminate between the various recommender system features and user variables.

Consecutively, the refined framework will be used in a large scale, long-term field trial, which will measure how the relative importance of the different user experience components (e.g. pragmatic and hedonic components) develops over time. The trial will be conducted in three countries (Germany, Spain and UK) in 2010. Each trial is based on the same underlying MyMedia software framework but the company branding, user interface design, system functionalities and content catalogues result in quite different front-ends. The aim is to compare the four Multimedia recommender systems on a set of evaluation criteria that is partially shared and partially tailored to the specific system. This allows for comparison across recommender systems, as well as the exploration of unique system features (e.g. social networking and recommender algorithms) or user group characteristics (e.g. age and cultural background).

The generic nature of the proposed evaluation framework allows it to be applied to any recommender system. Acknowledging the importance of user experience related factors that transcend algorithmic performance, it aspires to take a next step in the road towards usable and useful recommenders. We therefore encourage designers and researchers to use our framework as an integrative guideline for the design and evaluation of new generations of recommender systems.

Acknowledgement

This research is funded by the European Commission FP7 project MyMedia (www.mymediaproject.org) under the grant agreement no. 215006.

References

1. Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: User Acceptance of Computer Technology: A Comparison of two theoretical models. Management Science 35(8), 982–1003 (1989)
2. Hassenzahl, M.: User Experience (UX): Towards an experiential perspective on product quality. In: IHM 2008, Metz, France, September 2-5, pp. 11–15 (2008)
3. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems 22(1), 5–53 (2004)
4. Knijnenburg, B.P., Willemsen, M.C.: Understanding the Effect of Adaptive Preference Elicitation Methods on User Satisfaction of a Recommender System. In: ACM Conference on Recommender Systems 2009, New York (2009)
5. McNee, S.M., Riedl, J., Konstan, J.A.: Making recommendations better: An analytic model for human-recommender interaction. In: CHI 2006, Montreal Canada (2006)
6. McNee, S.M., Riedl, J., Konstan, J.A.: Accurate is not Always Good: How Accuracy Metrics have hurt Recommender Systems. In: CHI 2006, Montreal, Canada (2006)
7. Meesters, L., Marrow, P., Knijnenburg, B., Bouwhuis, D., Glancy, M.: Deliverable 1.5 End-user recommendation evaluation metrics. FP7 MyMedia project, `http://www.mymediaproject.org`
8. Spiekermann, S.: Online Information Search with Electronic Agents: Drivers, Impediments, and Privacy Issues. PhD Thesis, Humboldt University, Berlin (2001)
9. Xiao, B., Benbasat, I.: E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact. MIS Quarterly 31(1), 137–209 (2007)
10. Zins, A., Bauernfeind, U.: Explaining Online Purchase Planning Experiences with Recommender Websites. In: Information and Communication Technologies in Tourism 2005, Innsbruck, Austria, pp. 137–148 (2005)