

Recommender Systems

and the User-Centric Evaluation of Their Interfaces

Written by: Bart Knijnenburg : www.usabart.nl : bart@usabart.nl

Introduction

At the 2009 Recommender Systems conference, Francisco Martín (CEO of Strands Inc) argued that recommendation is not about algorithms. In fact, finding related products is relatively easy; the hard part is finding a good way to present recommendations and to gather preference feedback from the user. Consequently the user interface represents about 50% of the relevance of a recommender system, the algorithm only 5% [1].

Executive summary

Reflecting Martín's views, this document makes the following arguments:

- **Accuracy is not enough:** higher accuracy does not automatically mean better experience; there are other factors at play
- **Evaluation is comparison:** summative user-experience evaluation methods produce generalizable, statistically valid results
- An **integrated framework** provides the how and why of recommender systems user-experience
- A **pragmatic toolbox** simplifies this evaluation so that it can be conducted on a large scale in companies running live recommender systems

Accuracy is not enough

Despite this observation, most research in the field of Recommender Systems still focuses on the accuracy of recommendation algorithms. However, this research makes three implicit assumptions:

1. A higher accuracy brings the predicted topN recommendations closer to the real topN of the user
2. Recommendations close to the real topN are better
3. Better recommendations lead to a better user experience

I doubt that these premises hold. First of all, if an algorithm predicts a rating of 4.8 stars instead of 4.9, what difference does that make? In a sufficiently large set of items the first 100 recommendations will be good anyway, so an algorithm does not have to be that accurate. Secondly, are 5-star recommendations always the best? Users may instead want recommendations that are more varied or unpredictable than their topN. Finally, users may not want recommendations at all, but rather browse the system on their own.

Researchers occasionally discover that their most accurate algorithm results in the lowest user satisfaction, and vice versa [2, 3]. Moreover, researchers have shown that diversifying recommendations results in a higher satisfaction [4], that the preference elicitation method influences decision accuracy and intention to return [5], and that privacy concerns determine whether users are willing to disclose their personal information [6]. Such user-centric research, however, cannot be conducted on offline datasets. Instead, these tests need to be conducted with real users using real systems.

Evaluation is comparison

There are numerous ways to evaluate the user experience of a recommender system, and these can be categorized as either formative or summative, depending on the goal of the evaluation. In formative research, the goal is to improve a certain system A in terms of a certain quality X. The usual formative approach is to test system A qualitatively, focusing on quality X, looking for improvements on certain features P, Q and R that will increase quality X. Formative methods include think-aloud user studies, heuristic evaluation, and cognitive walkthrough. The goal of summative evaluation is to find out whether feature P causes quality X (regardless of the system that uses feature P). The usual summative approach is to test system A versus system B, where these systems only differ on feature P, and then measuring quality X to see if it differs between the two systems. Summative methods include A/B tests (field trials) and controlled experiments.

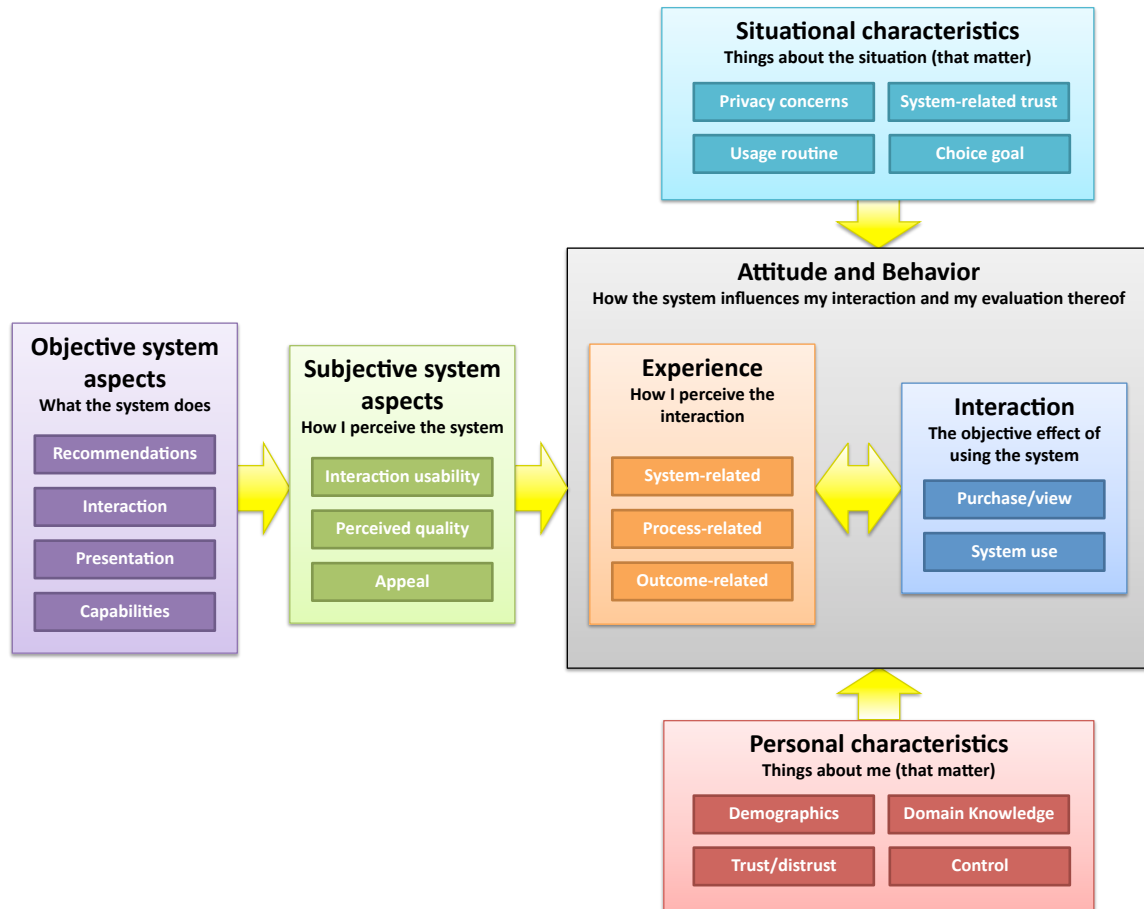
Formative evaluation is quicker and cheaper to conduct than summative evaluation, and the results are more straightforward. However, the method is less suitable for adaptive systems (including recommender systems), because it is hard to find out what exactly causes (problems with) the user experience. In summative evaluation you need to define hypotheses beforehand, you can only focus on a few aspects at a time, and the analysis is more complex. On the other hand, it is easier to test adaptive systems because you can single out the effect of specific features. Moreover, the results of summative evaluation are more generalizable and can be statistically validated.

Although I am fully aware of the advantages of formative research, and the invaluable role it can play in user experience evaluations, my personal preference goes summative evaluation. It reflects my background in research psychology, and it best represents the academic approach of generalizable, reproducible experiments. Below I introduce a user-centric evaluation framework and a pragmatic toolbox for online evaluation that can reduce the drawbacks of summative evaluation. Specifically:

- The user-centric evaluation framework can suggest a focus for the evaluation
- In the user-centric evaluation framework, different evaluations can be integrated
- The pragmatic toolbox reduces the resources required to deploy an evaluation
- The toolbox provides simple methods for data analysis

An integrated framework

Using the user-centric evaluation framework, one can evaluate the effect of a certain Objective System Aspect (OSA) on the user experience (EXP) and interaction (INT) of a recommender system, taking into account personal and situational characteristics (PC and SC).



The framework, depicted above, contains the following components:

Objective System Aspects (OSA): These are objective qualities of the recommender system, such as its design, its algorithm, the way it presents recommendations, and other system features.

Subjective System Aspects (SSA): These are the users' evaluations of the objective aspects. They include usability, quality and visual attractiveness. The subjective system aspects say more about the system than about the user.

User Experience (EXP): The user experience consists of self-relevant interpretations of the system aspects; it relates to the impact of the aspects on the user. Experience is multi-faceted: it can relate to the system, the process of using the system, or the outcome of using the system (the chosen items).

Personal and Situational Characteristics (PC and SC): These are characteristics of the user (PC) or the specific interaction instance (SC) that are beyond the control of the recommender system, but may influence the experience.

Interaction (INT): Interaction is measured as log-ins, clicks, views, and purchases with the system. These behaviors are objectively measurable (extracted from usage logs). They present the final step in the evaluation.

The evaluation framework links objective interaction (blue) to objective system aspects (purple) through a series of subjective constructs. There are three reasons for doing this:

1. The SSAs and EXP are likely to attenuate and qualify the effects of OSAs on INT. The direct effect may not be as strong as the mediated effect.
2. Some behaviors are not easily observed because they are rare, or occur over a longer period of time (e.g. adoption of the system). In this case, experience is a valid proxy of the behavior.
3. The mediation through SSA and EXP explains why users' behavior is different for different systems.

The framework can be used to conduct controlled experiments. OSAs can be "manipulated"; two or more systems can be created that differ only on the OSA. Participants in the experiment are then randomly assigned to one of the systems. SSAs, EXP, PCs and SCs can be measured using questionnaires. It is virtually impossible to measure a certain SSA or EXP with only a single question. Instead, I propose using 5-8 (positively and negatively phrased) statements to which users can agree or disagree on a 5- or 7-point scale. Interaction can be measured using data-logs. Using a combination of behavior and subjective evaluations is very useful: it grounds subjective data in "real" behavior, and helps the interpretation of objective measures. Finally, the hypothesized relationships between OSAs, SSAs, EXP, PCs, SCs, and INT can be tested using Structural Equation Modeling, a powerful statistical technique that is very suitable for this kind of data.

The framework has been validated in 4 field trials and 2 controlled experiments. For more details about the framework and examples of how to use it for the evaluation of specific experiments, please read [7, 8, 9].

A pragmatic toolbox

The proposed methodology to evaluate recommender systems is very thorough, but also rather complex and resource-intensive. Companies testing their deployed recommender systems do not want to bother their customers with endless questionnaires, and they do not have the resources to conduct intricate statistical analyses on the results of the evaluation. A pragmatic toolbox may provide a solution here.

The toolbox maintains the principles of controlled experimentation using manipulations, but simplifies the constructs and the analysis. Instead of conducting questionnaires with 5-8 items per construct, previous work based on the framework can suggest 1 or 2 items that

robustly measure the construct, or a behavioral measure (logged data) that can serve as a proxy for the construct. Instead of conducting a SEM analysis, our framework can suggest relations that should or could exist between constructs. These can then be tested using simple t-tests or correlations.

This toolbox could be deployed using Google Analytics, Google Website Optimizer, and Google Questionnaire (part of Google Spreadsheet) in order to create a fully automated recommender systems evaluation environment. This deployed toolbox could then be used by companies in full-scale evaluations of the user experience of their existing recommender systems.

References

- [1] Martin, F. J. 2009. Top 10 lessons learned developing, deploying, and operating real-world recommender systems. http://recsys.acm.org/2009/invited_talk_strands_martin.pdf.
- [2] McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J. 2002. On the Recommending of Citations for Research Papers. In: *Proceedings of CSCW'02*, pp. 116-125. ACM, New York
- [3] Torres, R., McNee, S.M., Abel, M., Konstan, J.A., Riedl, J. 2004. Enhancing Digital Libraries With TechLens+. In: *Proceedings of JCDL'04*, pp. 228-236. ACM, New York
- [4] Ziegler, C. N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving Recommendation Lists Through Topic Diversification. In: *Proceedings of WWW'05*, pp. 22-32. ACM, New York
- [5] Chen, L. Pu, P. 2009. Interaction design guidelines on critiquing-based recommender systems. In: *UMUAI 19*, pp. 167-206.
- [6] Kobsa, A., Teltzrow, M.: Contextualized Communication of Privacy Practices and Personalization Benefits: Impacts on Users' Data Sharing and Purchase Behavior. In: *Privacy Enhancing Technologies 3424*, pp. 329-343.
- [7] Knijnenburg, B. P., Meesters, L. M. J., Marrow, P., Bouwhuis, D. G. 2009. User-Centric Evaluation Framework for Multimedia Recommender Systems. In: *Proceedings of UCMedia'09, LNICST 40*, pp. 366-369. Springer-Verlag, Berlin Heidelberg
- [8] Knijnenburg, B. P., Willemsen, M. C., Hirtbach, S. 2010. Receiving recommendations and Providing Feedback: The User-Experience of a Recommender System. In: *Proceedings of UC-Web'10, LNBIP 61*, pp. 207-216. Springer-Verlag, Berlin Heidelberg
- [9] Knijnenburg, B. P., Willemsen, M. C. Forthcoming. An evaluation framework for explaining the why and how of the user experience of recommender systems. *Working draft obtainable from authors.*