Commentary on "A User-Centric Evaluation Framework for Recommender Systems"

The Study

The study, made by researchers Pearl Pu (EPFL), Li Chen (Hong Kong Baptist University) and Rong Hu (EPFL), consists of the creation of a questionnaire to evaluate the performance of virtually any recommender system. The two main contributions are the good quality of the questionnaire (measured by the authors with a survey) and the hierarchical structure of its constructs.

Commentary

The subjective nature of the evaluation of a recommender system makes me distrust any evaluation framework (at first glance, at least). I think the authors do a fine job of mitigating this – *i.e.* convince of trusting their framework – by validating their constructs and questionnaires in numerous ways while basing their research in previous work (both their own and others') on recommender systems' evaluation techniques.

The survey that validates their proposed questionnaire does not seem to be available any more at this site, which is the site provided in the paper's footer. This fact is unfortunate since access to the principal validation tool used by the authors is essential to grasp the study further. Although, understandably, the same website may not be working because many years have passed. Perhaps it has been moved, or the authors just assumed they would be contacted upon the need of their survey data.

Since I am not an expert in questionnaires' evaluation, I investigated the meaning of internal reliability, convergent validity, and inter-construct correlations. Regarding the first, I found an inconsistency. The authors claim that a value of 0.5 for Cronbach's alpha is an acceptable value when they write "Since some of the alpha values were under 0.5 (a value viewed as acceptable) […]". In contrast, George, D., & Mallery, P. (2003) in "SPSS for Windows step by step: A simple guide and reference" say it is poor and borderline unacceptable, as shown in the following table:

Cronbach's alpha	Internal consistency
$0.9 \leq \alpha$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

This inconsistency is essential to consider because the interpretation of the alpha value influences the validity of the whole questionnaire, which leads me to wish the authors had included a reference to support the claim that 0.5 is acceptable. Note that the constructs of their questionnaire achieve alphas between 0.636 and 0.855, values spanning the range from "**questionable"** to "good" in the table above.

Overall, however, I believe the study represents an admirable effort (and, to my knowledge, a successful one) to test recommender systems from a user's perspective. It rightly incorporates aspects and contributions from fields such as psychology and sociology, which concern human (user) behaviour. Also, they subject their proposals, namely their questionnaire and the hierarchical relations of its constructs, to intense scrutiny, which is often lacking or poorly performed in some studies.