Tag-Enhanced Collaborative Filtering for Increasing Transparency and Interactive Control

Tim Donkers
University of Duisburg-Essen
Duisburg, Germany
tim.donkers@uni-due.de

Benedikt Loepp
University of Duisburg-Essen
Duisburg, Germany
benedikt.loepp@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT
To increase transparency and interactive control in Recommender Systems, we extended the Matrix Factorization technique widely used in Collaborative Filtering by learning an integrated model of user-generated tags and latent factors derived from user ratings. Our approach enables users to manipulate their preference profile expressed implicitly in the (intransparent) factor space through explicitly presented tags. Furthermore, it seems helpful in cold-start situations since user preferences can be elicited via meaningful tags instead of ratings. We evaluate this approach and present a user study that to our knowledge is the most extensive empirical study of tag-enhanced recommending to date. Among other findings, we obtained promising results in terms of recommendation quality and perceived transparency, as well as regarding user experience, which we analyzed by Structural Equation Modeling.

Keywords
Recommender Systems; Interactive Recommending; Matrix Factorization; Tags; Human Factors; User Experience

1. INTRODUCTION
Letting users influence the recommendation process and making it more comprehensible is increasingly considered an important goal in *Recommender Systems* (RS) research [17, 27, 15]. Interactive RS have been proposed that use, for instance, user-provided tags for eliciting preferences [30]. This has the advantage of relying on concepts that are meaningful to users without requiring explicit item descriptions, thus being promising for improving user control and comprehension [28]. However, tag-based RS in general (e.g. [28, 30]), and, specifically attempts to increase interactivity (e.g. [4, 2]), have mostly been developed independently of established *Collaborative Filtering* (CF) methods, and can therefore not benefit from existing long-term user profiles based on rating data or implicit feedback. Moreover, the availability of precise and efficient model-based CF algorithms such as the widely used *Matrix Factorization* (MF) [18] is usually not exploited. What is lacking, thus, are combinations of the accuracy-related benefits of model-based RS with the easy-to-understand semantics of tags.

We recently proposed an interactive recommending approach that integrates latent factors automatically derived by MF with tags users provided for the items [5]. In contrast to other approaches that enhance latent factor models with further data [13, 9, 22, 29, 8], we utilize the additional information to also allow users to interactively express their preferences and control the recommendation process through selecting and weighting tags, thus indirectly determining their profile in the factor space. Besides, the tags serve as a means to elicit preferences in cold-start situations without requiring the user to rate items. Following offline experiments [5], we now present a user study with a prototype system to further evaluate our approach. To our knowledge, it is the most extensive empirical study of a tag-enhanced RS to date. Among several promising findings, e.g. regarding choice difficulty and interaction process, it shows that integrating tags into MF also increases perceived recommendation quality, which previously has only been observed offline. To analyze the user experience we used *Structural Equation Modeling* (SEM) [23]. Still rarely used in RS research [15], this method gave us interesting insights into user behavior when tag-based interaction is offered in a RS, and emphasizes the value of the increased level of transparency introduced by our approach.

2. RELATED WORK AND BACKGROUND
Interactive recommenders (e.g. [30, 4, 2, 20]) are especially useful in cold-start situations when no historical data is available for new users or when a user does not want an existing profile to be applied. This common issue has been addressed in CF in several ways, but attempts to increase interactivity are overall typically independent of CF: Various approaches have been proposed, but they usually rely on their own concepts to recommend items instead of building on the benefits of established model-based CF techniques. Thus, even when available, previously given ratings or past browsing behavior cannot be considered. Overall, the availability of precise and efficient algorithms such as MF is typically not exploited by interactive RS. Latent factor models, in particular MF, have in turn been improved primarily with respect to objective accuracy metrics [17, 15], for instance, by complementing ratings with further data. The additional information used may be rather generic, such as implicit feedback or temporal effects [18], but also more specific, predefined metadata are taken into account [9]. Other approaches integrate the models with contextual information [13] or topics inferred from semantically analyzed product reviews [22]. In contrast, only a few rely on user-provided information such as tags [29, 8]. There are indeed recommending approaches that primarily use tags [28, 12, 25], but apart from e.g. *MovieTuner* [30], these tag-based RS are not particularly aimed at giving users more control. Developed independently of model-based CF techniques, they also cannot benefit from the algorithms' maturity and the availability of explicit or implicit user rating feedback.

The range of methods for integrating further data into CF is very broad. When using common SVD-like MF [18], constraints or regularization terms may be added when training the model [18, 22, 29, 8]. However, the latent factors then exhibit no interpretable association with the additional information as this is calculated into the factor values. Thus, the relationship between data and factors, and consequently items, cannot be accessed by users. In contrast, in [9], a content-related association with the factors is explicitly established: By proposing a regression-constrained formulation, they are considered as functions of content attributes. In our previous work [5], we initially followed this approach by integrating item-specific tag relevance information, but then also derived user-tag relevance scores as well as tag-factor relations. With $\mathbf{P}\in\mathbb{R}^{|U|\times|F|}$ and $\mathbf{Q}\in\mathbb{R}^{|I|\times|F|}$ being the user/item-factor matrices, this leads to:

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q}^{\mathrm{T}} = {}^{\mathbf{u}}\mathbf{A}\,\mathbf{\Lambda}\,{}^{\mathbf{i}}\mathbf{A}^{\mathrm{T}}$$

where ${}^{\mathbf{u}}\mathbf{A}\in\mathbb{R}^{|U|\times|T|}$ describes how strongly each user relates to each tag, ${}^{\mathbf{i}}\mathbf{A}\in\mathbb{R}^{|I|\times|T|}$ is the equivalent on item side, and $\mathbf{\Lambda}\in\mathbb{R}^{|T|\times|T|}$ contains the latent factor information. This method proposed in [5] thus gives us the opportunity to access the previously abstract user/item-factor vectors in a much more comprehensible way: Based on the model learned, user profiles now comprise information related to both tags and latent factors. As the tag concept is easily understood, this allows us to let users actively adjust their user profile. Therefore, we define a weight vector $w_u\in[0,1]^{|T|}$ to hold the user feedback regarding the tags (where 0 means no and 1 maximal interest), which is added to a_u for calculating recommendations:

$$\tilde{r}_{ui} = (a_u + w_u)\,\mathbf{\Lambda}\,a_i^{\mathrm{T}}$$

Beyond that, latent factor models have only rarely been exploited for purposes other than improving algorithmic performance. Exceptions comprise visualizations [10, 24] or choice-based preference elicitation methods [21, 11]. Still, the derived factors are overall hard to explain and it is particularly difficult from a system-perspective to relate them to an intelligible meaning [18]. Thus, users lack a deeper understanding of the recommendations and can typically not be provided with interactive control. While such aspects related to user experience are increasingly considered important for RS research, only few evaluations go beyond measuring accuracy in offline experiments [17, 27, 15]. Tag-enhanced RS have not been extensively analyzed by user studies, and especially integrating additional data into latent factor models has not yet been examined in terms of its actual influence on users. To evaluate user experience, the model proposed in [15] may serve as an important means that explains how subjective system aspects (e.g. perceived quality or effort) mediate the influence of objective system aspects (e.g. differences in recommender algorithms). Although considered particularly useful, advanced methods such as SEM are however only rarely used in RS research [15]. Exceptions have investigated, for instance, effects of objective system aspects on user perception of results [6], influence of choice-based preference elicitation compared to conventional ratings [11], or how the number of recommended items affects choice difficulty and satisfaction [1].

3. EMPIRICAL USER STUDY

To demonstrate and to evaluate our tag-enhanced recommending approach proposed in [5], we developed a web-based movie RS and conducted a user study. We used the Stochastic Gradient Descent MF algorithm from the Apache Mahout library as a baseline, and extended this algorithm (in same configuration) according to our method considering a limited number of 25 most popular tags as additional training data. We also implemented online-updating of factor vectors. As datasource for items and associated ratings and tags, we created an intersection of the well-known MovieLens 10M dataset and the MovieLens Tag Genome dataset. Reducing these datasets to those movies included in both left us with 8 429 movies, 9 964 745 ratings and 9 507 912 tag relevance scores.

For comparing our approach with conventional CF, we implemented two versions of the system: one used the standard MF algorithm, the other our tag-enhanced method. In the standard MF version, the top-10 recommended movies were displayed together with their movie posters and some metadata. Users could only rate the items recommended and explicitly search further titles in order to rate them as well. Upon rating an item, the result set was updated immediately. In the version based on our approach, users could additionally select tags and change their weight. For a screenshot and a more detailed description, please refer to [5].

3.1 Goals

We were especially interested in evaluating user experience as well as subjective perception of recommendation quality, transparency, and in particular, the preference elicitation in our system and its interactive features. We hypothesized that including tags would lead to better recommendations in terms of perceived quality, and would also increase transparency of the results, especially in cold-start situations. We also assumed that users would prefer tag-based interaction while the perceived effort would be acceptable despite the increased level of interactivity offered by our approach.

3.2 Method

Participants: We recruited 46 participants (33 female) with an average age of 22.89 (*SD*=6.88), most of them students (85 %). The study was designed as an experiment under controlled conditions.

Questionnaire/Log data: Participants had to fill in a questionnaire that was primarily based on the pragmatic evaluation procedure for RS described in [16], containing items regarding, among others, recommendation quality and usage effort. This framework based on [15] is reduced to stable operationalizations of the subjective constructs and, after repeated validation, appears to measure user experience in RS reasonably well with a limited number of questionnaire items [16]. In addition, we used items from [26] to assess recommendation transparency and interface adequacy, as well as self-generated items to ask which system version participants prefer. Further, we applied the *System Usability Scale* (SUS [3]) and *User Experience Questionnaire* (UEQ [19]), and gathered data about demographics, participants' interest in movies and their familiarity with this domain. Apart from UEQ (7-point bipolar scale), all items were assessed on a positive 5-point Likert-scale (1–5). We also logged users' interaction behavior and measured task times.

Procedure: First, participants were asked to complete two preliminary tasks in counter-balanced order that served to elicit an initial set of preferences. In one task, participants were asked to rate 10 out of the 30 most popular movies. Items were shown in random order and could be skipped when unknown. In the other task, participants should select 3 tags they liked out of the 20 most popular ones (also shown in random order), which are then used to initialize a meaningful user-tag vector a_u. Next, based on the two system versions implementing standard MF and our tag-enhanced approach, respectively, we assigned the participants in counter-balanced order to three different conditions in a within-subject design:

1. *Standard MF:* Standard MF with initial recommendations based on the 10 user ratings. The only interaction possible was to rate more items.
2. *TMF-Rating:* Our tag-enhanced MF with initial recommendations based on the 10 user ratings. Users could interactively select and weight tags, and also rate more items.
3. *TMF-Tag:* Our tag-enhanced MF with initial recommendations based on the 3 selected tags. User interaction was similar to the previous condition.

In each condition, participants were initially shown the top-6 recommendations obtained with the respective algorithm. First, they were asked to choose one movie from the six recommended ones they would actually like to watch. Second, they rated their satisfaction with each recommendation on a 5-point Likert-scale (1–5). Third, they filled in the questionnaire described above. Next, participants were asked to interact with the current system version to further refine recommendations and to receive a result set that better matched their personal interests. After participants finished interaction at their own discretion, they were again presented with the (now adjusted) top-6 recommendations. Again, they had to select one movie out of them, rate how satisfying each recommendation was, and fill in a questionnaire (which was now complemented with questions regarding the interaction process). For each condition, the respective variables were thus assessed at two different points in time, before and after the corresponding interaction phase.

3.3 Results

Participants reported that they liked movies a lot (M=4.22, SD=0.63) while having average knowledge about movies in general (M=3.07, SD=0.80) as well as about newer movies (M=2.93, SD=0.98). We conducted two-way repeated measures ANOVAs to compare the effect of condition and point in time on the dependent variables. For the three conditions, marginal mean values and standard errors are presented in Table 1.

Table 1. Results for the different conditions.

	MF		TMF-Rating		TMF-Tag	
	M	*SE*	*M*	*SE*	*M*	*SE*
Perc. Rec. Quality	3.16	0.11	3.31	0.13	3.65	0.10
Mean Item Rating	3.11	0.10	3.29	0.11	3.55	0.10
Choice Satisfaction	4.00	0.10	4.10	0.13	4.35	0.09
Choice Difficulty	34.50	3.10	27.80	2.62	28.37	2.32
Transparency	3.20	0.15	3.41	0.15	3.73	0.13

Perceived Recommendation Quality: Concerning subjective quality, there was a statistically significant effect with α=0.05 for condition, $F(2,90)$=7.40, p<.001. Post hoc comparisons using Bonferroni correction indicate that the mean score for TMF-Tag was significantly higher than for both, TMF-Rating, p=.028, and standard MF, p<.001. However, there was no significant difference regarding perceived quality of recommendations before and after the interaction phase, $F(1,45)$=0.02, p=.904.

Mean Item Rating: We found similar differences between the conditions with regard to the individual satisfaction participants stated for each recommended item, $F(2,90)$=11.19, p=.001. Again, TMF-Tag received significantly higher ratings than TMF-Rating, p=.025, and standard MF, p<.001.

Choice Satisfaction: Regarding satisfaction with the movie participants finally selected from the set of recommendations, we also found statistical evidence for differences between the conditions, $F(2,90)$=4.72, p=.011. Post hoc Bonferroni tests indicate that the mean score for TMF-Tag was significantly higher than for standard MF, p=.009. No differences were found between TMF-Rating and other conditions. Furthermore, the mean values before and after the interaction tasks are statistically discriminable, $F(1,45)$=5.07, p=.029. Before interaction (M=4.28, SE=0.10) users were more confident with their selected movies than afterwards (M=4.02, SE=0.11). Since the interaction term of condition and point in time was not significant, we deduce that this applies to all conditions.

Choice Difficulty: With respect to objective difficulty to decide, operationalized as the total time participants spent for choosing one movie they would actually like to watch from the shown recommendations, the within-subjects main effect yielded significant differences for condition, $F(2,90)$=6.42, p=0.02. Post hoc comparisons denote that users took more time to decide in the standard MF condition compared to TMF-Rating, p=.012, and TMF-Tag, p=.027. Additionally, users tended to decide more quickly after they interacted with the system, $F(1,45)$=29.23, p<.001.

Transparency: We also found a significant effect of condition on transparency, $F(2,90)$=6.22, p=.003. Results from standard MF were perceived less transparent than from TMF-Tag, p=.003. No differences were found between TMF-Rating and other conditions.

Effort and Usability: The version allowing for interaction via tags was assessed significantly ($t(45)$=4.15, p<.001) better (M=3.76, SD=1.02) than the other (M=2.83, SD=1.00). Without tags, participants spent 165.54 sec (SD=114.64) for the entire interaction task, in the two conditions with tags, they needed on average 209.68 sec (SD=103.26). Although interaction phases were thus significantly longer ($t(45)$=−2.43, p=.019), perceived interaction effort was not higher: a one-way ANOVA yielded no significant effect for condition, $F(2,90)$=1.40, p=.253. Also, the usability was rated as "good" with a SUS-score of 78 and values between 0.95 and 1.96 on the different scales of the UEQ. In particular, the subscale for transparency yielded an excellent score (M=1.96) and efficiency was rated above average (M=1.16), which corresponds to the very promising assessment of interface adequacy (M=4.13, SD=0.48).

Structural Equation Modeling: Using SEM we further analyzed the questionnaire data to investigate the effects of varying recommender algorithm (*Standard MF* vs. *Tag-enhanced MF*) and method for eliciting initial preferences (*Ratings* vs. *Tags*) on user experience and interaction behavior. We were especially interested in differences between the three conditions in cold-start where the system must deal with a high level of uncertainty when presenting the first recommendations. We also considered personal characteristics to deduce assumptions about how different dispositions may influence those relations. Following [15], we define algorithms and preference elicitation methods as *Objective System Aspects* (OSA) that cannot be influenced by the user. *Perceived Rec. Quality* and *Transparency* are seen as *Subjective System Aspects* (SSA), which represent the user's perception of OSA. SSA are conceived as mediating variables between OSA and user experience [15]. User experience may be substantially influenced by using different algorithms and preference elicitation methods (see, e.g. [14, 4, 6, 15, 7]). We assume that user experience is affected by changes with respect to *Perceived Rec. Quality* and *Transparency* when a novel means for eliciting initial preferences is used, i.e. selecting tags according to our approach. We included *Choice Satisfaction* as an indicator of the user's *Experience* (EXP). The user's *Interaction Behavior* (INT) is also influenced by SSA. We therefore complement the more general *Perceived Rec. Quality* by capturing the specific feedback regarding each recommended item, i.e. the *Mean Item Rating*. Finally, we in line with the underlying framework assume that certain *Personal Characteristics* (PC) such as *Domain Knowledge* and *Trust in Technology* have an impact on attitude and behavior concerning the varied system aspects.

We set up the theoretical model shown in Figure 1 that yielded a good fit with the data ($\chi^2(12)$=13.669, p=.322, CFI=.995, TLI=.989, $RMSEA$=.032). It explains a large amount of variance regarding our dependent variables *Choice Satisfaction* (R^2=.401), *Mean Item Rating* (R^2=.693) and *Perceived Rec. Quality* (R^2=.523), and also a reasonable proportion with respect to *Transparency* (R^2=.234). Direct effects of the two different algorithms used in the three conditions were not significant for any dependent variable or the mediator. Thus, the algorithms (*Standard MF* vs. *Tag-enhanced MF*) were eventually not considered in our model. In contrast, the variation of the preference elicitation method (*Ratings* vs. *Tags*) seems to account for a significant explanation of *Transparency*. While *Domain Knowledge* as one of the personal characteristics shows a meaningful influence only on *Transparency*, *Trust in Technology* also influences *Choice Satisfaction* and *Mean Item Rating*. Further analysis shows that *Transparency* seems to be a substantial causal factor for *Perceived Rec. Quality*, which is an overall subjective assessment that in turn acts as a complete mediator for the effects on our dependent variables, i.e. the more specific *Choice Satisfaction* and *Mean Item Rating*. In particular, this route appears to completely mediate the otherwise significant predictive power of the different methods to elicit initial preferences (*Ratings* vs. *Tags*).

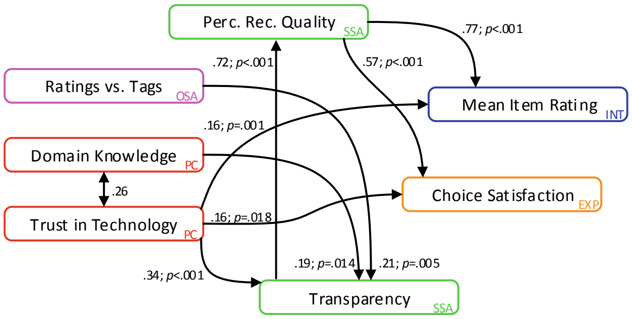

Figure 1. Path model for comparing the influence of preference elicitation via ratings or tags. On the edges, standardized regression weights and p-values are displayed.

3.4 Discussion

In general, including additional content information into MF seems to be beneficial in terms of objective recommendation quality. We observed this for our approach [5], thereby validating results of offline experiments performed by several others, e.g. [13, 9, 29, 22, 8]. However, by conducting a user study we could for the first time confirm that this finding also applies to the users' subjective perception. TMF-Tag received significantly higher scores with respect to perceived recommendation quality, satisfaction with the chosen movie, and transparency. Significant differences between conditions before the interaction phases (with tag-enhanced MF being superior) further suggest that the few interaction steps performed at the beginning to elicit preferences by selecting a small number of tags are already sufficient to improve user experience, in particular perceived quality and transparency. Regarding choice difficulty, condition and point in time both account for significant effects. The latter was to be expected as users may already have decided for an item during interaction, and therefore needed less time to settle on a recommended movie. However, it is particularly interesting that with standard MF, participants needed significantly longer to select a movie than in the tag-enhanced conditions. They further perceived recommendations to be significantly more transparent with TMF-Tag—also before the interaction, without knowing that the results were just based on the initially selected tags. Our tag-based preference elicitation approach thus seems to help users also implicitly when judging recommendations.

Because of these findings, we further examined the role of transparency in context of generating satisfying recommendations, particularly in cold-start situations, by using SEM. As indicated by our model, selecting tags instead of rating items to elicit initial preferences significantly improves transparency. We therefore deduce that tags import semantics into the result set which are more natural to understand by users than deriving a meaning from recommendations based on numerical ratings. Thus, our tag-enhanced recommending approach seems to lead to more comprehensible results. In general, increasing transparency seems to positively influence user satisfaction with recommendations. The high standardized regression weight of .72 supports that transparency is a substantial predictor. Consequently, the significant influence of preference elicitation method on transparency emphasizes that our approach is a promising means to alleviate the cold-start problem.

The fact that only varying the algorithm yielded no significant differences in recommendation quality is generally in line with recent research stating that different or objectively more accurate recommenders do not necessarily produce better results from a subjective perspective [17, 27, 6]. Instead, the entire result set should express some kind of inner consistency, which in case of our proposed method is reached through relating latent factors learned by MF with user-provided tags. While even increasing objective accuracy [5], our approach to use tags for eliciting preferences thus makes recommendations more transparent and thereby in fact also improves perceived recommendation quality. The recommendations then seem to refer to each other implied by the easy-to-understand semantics of tags. Conversely, although it may achieve high accuracy scores, a list of items detached from such a meaningful superordinate context might not be as satisfactory for the user.

Highly significant regression weights suggest that recommendation quality is the main predictor for choice satisfaction and users' individual rating feedback for recommended items. However, also domain knowledge, for instance, may increase users' satisfaction with the results as it helps to better comprehend why certain items were recommended. By increasing transparency, our approach thus seems to be especially useful for users with little domain knowledge. The influence of trust in technology on the dependent variables is in contrast not fully mediated via transparency. This is another indicator for the importance of aspects that go beyond recommendations themselves: Concerning the satisfaction with chosen item as well as mean item rating, it suggests that some personal characteristics might alter the way perceived recommendation quality is translated into numerical ratings. As a result, datasets comprising user feedback in form of ratings may suffer from nonsystematic deviations, i.e. users whose trust in technology is low would provide lower ratings in a more technically-oriented system. This, in turn, is another argument for using more natural ways than ratings to interact with CF recommenders.

Finally, regarding interaction, participants assessed our system's usability very positively. They preferred the tag-enhanced version, which might be a reason why they spent more time using it. Although the richer possibilities for interaction may also contribute to this finding, perceived interaction effort did not differ. Overall, our interactive recommending approach based on tags thus seems to be of value for providing users with more control over the recommendation process and for improving its transparency.

4. CONCLUSIONS AND OUTLOOK

Our user study confirmed that additional content information can be used in conjunction with MF not only to increase accuracy, but at the same time also improves perceived recommendation quality and transparency. Furthermore, users were more satisfied with the chosen movie while it was easier to settle for an item. Interestingly, besides the fact that users liked the interaction via tags generally more than just rating items, tag-enhanced MF yields particularly promising results when preferences were elicited initially by tags. Thus, our approach seems useful to interactively adapt results when a rating-based profile is already available as well as to set up a new profile in cold-start situations since a small number of selected tags leads to a user profile at least as good as when rating a larger number of items up front. Using SEM, we further analyzed these findings, focusing on the role of this new method to elicit initial preferences and its positive influence on the aforementioned aspects. In future work, we plan to exploit tags as well as other content-related or contextual data more extensively, for instance, to explain user profiles and to establish even richer interaction possibilities in MF recommenders. Finally, we are interested in comparing our system with other tag-based RS and in adapting it to different domains.

5. REFERENCES

[1] Bollen, D., Knijnenburg, B. P., Willemsen, M. C., and Graus, M. P. 2010. Understanding choice overload in recommender systems. In *Proc. RecSys '10*. ACM, 63–70.

[2] Bostandjiev, S., O'Donovan, J., and Höllerer, T. 2012. TasteWeights: A visual interactive hybrid recommender system. In *Proc. RecSys '12*. ACM, 35–42.

[3] Brooke. J. 1996. SUS – A quick and dirty usability scale. In *Usability Evaluation in Industry*, Taylor & Francis, 189–194.

[4] Chen, L. and Pu, P. 2012. Critiquing-based recommenders: Survey and emerging trends. *User Mod. User-Adap. 22*, 1-2, 125–150.

[5] Donkers, T., Loepp, B., and Ziegler, J. 2015. Merging latent factors and tags to increase interactive control of recommendations. In *Poster Proc. RecSys '15*.

[6] Ekstrand, M. D., Harper, F. M., Willemsen, M. C., and Konstan, J. A. 2014. User perception of differences in recommender algorithms. In *Proc. RecSys '14*. ACM, 161–168.

[7] Ekstrand, M. D., Kluver, D., Harper, F. M., and Konstan, J. A. 2015. Letting users choose recommender algorithms: An experimental study. In *Proc. RecSys '15*. ACM, 11–18.

[8] Fernández-Tobías, I. and Cantador, I. 2014. Exploiting social tags in matrix factorization models for cross-domain collaborative filtering. In *Proc. CBRecSys '14*, 34–41.

[9] Forbes, P. and Zhu, M. 2011. Content-boosted matrix factorization for recommender systems: Experiments with recipe recommendation. In *Proc. RecSys '11*. ACM, 261–264.

[10] Gansner, E., Hu, Y., Kobourov, S., and Volinsky, C. 2009. Putting recommendations on the map: Visualizing clusters and relations. In *Proc. RecSys '09*. ACM, 345–348.

[11] Graus, M. P. and Willemsen, M. C. 2015. Improving the user experience during cold start through choice-based preference elicitation. In *Proc. RecSys '15*. ACM, 273–276.

[12] Guan, Z., Wang, C., Bu, J., Chen, C., Yang, K., Cai, D., and He, X. 2010. Document recommendation in social tagging services. In *Proc. WWW '10*. ACM, 391–400.

[13] Karatzoglou, A., Amatriain, X., Baltrunas, L., and Oliver, N. 2010. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proc. RecSys '10*. ACM, 79–86.

[14] Knijnenburg, B. P. and Willemsen, M. C. 2010. The effect of preference elicitation methods on the user experience of a recommender system. In *Ext. Abstr. CHI '10*. ACM, 3457–3462.

[15] Knijnenburg, B. P. and Willemsen, M. C. 2015. Evaluating recommender systems with user experiments. In *Recommender Systems Handbook*. Springer US, 309–352.

[16] Knijnenburg, B. P., Willemsen, M. C., and Kobsa, A. 2011. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proc. RecSys '11*. ACM, 321–324.

[17] Konstan, J. A. and Riedl, J. 2012. Recommender systems: From algorithms to user experience. *User Mod. User-Adap. 22*, 1-2, 101–123.

[18] Koren, Y., Bell, R. M., and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *IEEE Computer 42*, 8, 30–37.

[19] Laugwitz, B., Held, T., and Schrepp, M. 2008. Construction and evaluation of a user experience questionnaire. In *Proc. USAB '08*. Springer, 63–76.

[20] Loepp, B., Herrmanny, K., and Ziegler, J. 2015. Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques. In *Proc. CHI '15*. ACM, 975–984.

[21] Loepp, B., Hussein, T., and Ziegler, J. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proc. CHI '14*. ACM, 3085–3094.

[22] McAuley, J. and Leskovec, J. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proc. RecSys '13*. ACM, 165–172.

[23] Muthén, B. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika 49*, 1, 115–132.

[24] Németh, B., Takács, G., Pilászy, I., and Tikk, D. 2013. Visualization of movie features in collaborative filtering. In *Proc. SoMeT '13*, 229–233.

[25] Nguyen, T. T. and Riedl, J. 2013. Predicting users' preference from tag relevance. In *Proc. UMAP '13*. Springer, 274–280.

[26] Pu, P., Chen, L., and Hu, R. 2011. A user-centric evaluation framework for recommender systems. In *Proc. RecSys '11*. ACM, 157–164.

[27] Pu, P., Chen, L., and Hu, R. 2012. Evaluating recommender systems from the user's perspective: Survey of the state of the art. *User Mod. User-Adap. 22*, 4-5, 317–355.

[28] Sen, S., Vig, J., and Riedl, J. 2009. Tagommenders: Connecting users to items through tags. In *Proc. WWW '09*. ACM, 671–680.

[29] Shi, Y., Larson, M., and Hanjalic, A. 2013. Mining contextual movie similarity with matrix factorization for context-aware recommendation. *ACM Trans. Intell. Syst. Technol. 4*, 1, 16:1–16:19.

[30] Vig, J., Sen, S., and Riedl, J. 2011. Navigating the tag genome. In *Proc. IUI '11*. ACM, 93–102.