Interactive Recommending with Tag-Enhanced Matrix Factorization (*TagMF*)[☆]

Benedikt Loepp*, Tim Donkers, Timm Kleemann, Jürgen Ziegler

University of Duisburg-Essen, Duisburg, Germany

Abstract

We introduce *TagMF*, a model-based Collaborative Filtering method that aims at increasing transparency and offering richer interaction possibilities in current Recommender Systems. Model-based Collaborative Filtering is currently the most popular method that predominantly uses Matrix Factorization: This technique achieves high accuracy in recommending interesting items to individual users by learning latent factors from implicit feedback or ratings the community of users provided for the items. However, the model learned and the resulting recommendations can neither be explained, nor can users be enabled to influence the recommendation process except by rating (more) items. In *TagMF*, we enhance a latent factor model with additional content information, specifically tags users provided for the items. The main contributions of our method are to use this integrated model to elucidate the hidden semantics of the latent factors and to let users interactively control recommendations by changing the influence of the factors through easily comprehensible tags: Users can express their interests, interactively manipulate results, and critique recommended items—at cold-start when no historical data is yet available for a new user, as well as in case a long-term profile representing the current user's preferences already exists.

To validate our method, we performed offline experiments and conducted two empirical user studies where we compared a recommender that employs *TagMF* against two established baselines, standard Matrix Factorization based on ratings, and a purely tag-based interactive approach. This user-centric evaluation confirmed that enhancing a model-based method with additional information positively affects perceived recommendation quality. Moreover, recommendations were considered more transparent and users were more satisfied with their final choice. Overall, learning an integrated model and implementing the interactive features that become possible as an extension to contemporary systems with *TagMF* appears beneficial for the subjective assessment of several system aspects, the level of control users are able to exert over the recommendation process, as well as user experience in general.

Keywords: Recommender Systems, Collaborative Filtering, Interactive Recommending, Matrix Factorization, Tags, Empirical Studies, Human Factors, User Experience, User Interfaces, User Profiles

1. Introduction

Recommender Systems (RS) based on *Collaborative Filtering* (CF) have been shown to be effective means for leveraging the "wisdom of the crowd" to identify items that are potentially of interest to a user. They support users in finding items that match their personal preferences from very large sets of items, such as, for instance, consumer goods, documents, or movies [1, 2]. From an information provider's perspective, a major advantage of CF recommenders lies in the fact that only feedback the community of users provided for the items—explicitly expressed via ratings or implicitly acquired through user actions—is required as input data [3]. Considerable advances have been made in recent years with respect to the objective performance of CF systems as measured by common accuracy metrics in retrospective offline experiments [4]. However, it has been observed that high offline recommendation accuracy (i.e. accurately predicting which items should be recommended to a user) does not necessarily lead to a commensurate level of user satisfaction [5, 6, 7]. Since CF algorithms are considered already quite mature, the small incremental improvements that still seem possible with respect to algorithmic precision are thus not likely to be particularly beneficial for users. Consequently, other evaluation metrics have been discussed to assess the quality of recommendation sets, for example, diversity, novelty, and serendipity [8, 9]. Beyond that, one important aspect that may contribute to actual user satisfaction is the degree of control users have over the systems [6, 10]. Yet, from a user's perspective, the ways to influence the generation of recommendations in today's automated RS such as the ones used by Amazon [11] or Netflix [12] are mostly very limited. Usually, the only means to actively influence the results is to provide explicit feedback about the items' relevance, i.e. rating or re-rating single items. Among others, this raises the risk of users being stuck in a "filter bubble" [13] as the recommendations are increas-

 DOI: 10.1016/j.ijhcs.2018.05.002.

*Corresponding author

Email addresses: benedikt.loepp@uni-due.de (Benedikt Loepp), tim.donkers@uni-due.de (Tim Donkers), timm.kleemann@uni-due.de (Timm Kleemann), juergen.ziegler@uni-due.de (Jürgen Ziegler)

ingly constrained to items similar to those the current user has rated positively in the past. This well-known effect makes it difficult to become aware of hidden alternatives, to explore new and diverse areas of potential interest, and to adapt the results towards situational needs and goals [13, 14]. A further problem can be seen in the general lack of transparency of contemporary CF recommenders [5, 7]. The methods prevalently used infer abstract models from the original input data, making it difficult for users to understand the profile that represents their preferences, and consequently why certain items are recommended. This, in turn, may reduce user trust in the system as well as acceptance of suggested items [5, 15]. Hence, adding more interactivity to the system and letting users influence the recommendation process as well as making it more comprehensible is increasingly considered an important goal in RS research [5, 6, 7, 16, 10].

Only more recently, such aspects related to user experience of RS have begun to attract more attention [6, 17]. In this line of research, interactive recommenders have been proposed that use, for instance, metadata such as user-generated tags to calculate recommendations and to offer users additional interaction mechanisms [18]. Using tags has the advantage of relying on concepts that are meaningful to users without requiring explicit item descriptions. Consequently, eliciting preferences via tags has been shown to bear the potential for improving user control and comprehension [19]. However, tag-based RS [e.g. 19, 20, 18] have typically been developed independently of established CF methods. For this reason, such systems cannot benefit from existing long-term preference profiles based on implicit feedback or rating data. The same applies for most of the recommending approaches that aim at increasing interactivity in general [e.g. 21, 22, 23]. In particular, they usually do not exploit model-based CF techniques such as the widely used *Matrix Factorization* (MF) [24, 25], which is known for efficiency and has been shown to achieve high offline accuracy. On the other hand, models as derived by MF have only rarely been applied for purposes other than improving recommendation effectiveness or algorithmic performance. What is lacking, therefore, are methods that combine the accuracy-related benefits of model-based CF with the easy-to-understand semantics of user-generated tags.

In this paper, we introduce *Tag-Enhanced Matrix Factorization* (*TagMF*), a novel recommendation method that enhances a MF model with tags that users provided for the items, and propose several possible applications of *TagMF* for realizing interactive recommenders by extending conventional model-based CF systems. MF models represent both users and items in a joint latent factor space [24]. Since latent factors are usually learned by statistically analyzing historical implicit feedback or rating data, the semantics of these factors are hidden, yet are generally considered to relate to real-world concepts [24, 26]. For instance, the factors may describe more or less obvious characteristics such as the "amount of action" a user appreciates or the "degree of black humor" in a movie. Once the factors have been learned, latent factor models allow to accurately predict ratings for items the user has not yet seen, or to establish a ranking among them. Several approaches already exist that employ additional information such as context data, predefined or user-generated content-related metadata, or topics and opinions inferred from user reviews [e.g. 27, 28, 29, 30, 31, 32, 33, 34]. The respective methods have been applied with the goal of further improving model quality, and as a consequence offline recommendation accuracy (i.e. how well the predictions match implicit feedback or ratings provided in the past), not for exposing the additional information at the user interface. Accordingly, there are currently also no user studies that show the benefits of enhancing a model-based CF recommender with additional content information. With *TagMF*, we contribute to the state of research by answering the following research questions:

RQ1: How can additional information be used in model-based CF systems for ...
- **a)** eliciting preferences in cold-start situations without requiring the user to rate items?
- **b)** manipulating recommendations resulting from an existing user profile?
- **c)** critiquing a recommended item while considering the user's long-term interests?
- **d)** explaining an existing preference profile?

RQ2: How does additional information affect subjective system aspects such as perceived recommendation quality and user experience when compared to ...
- **a)** an automated recommender based on ratings?
- **b)** an interactive recommender based on tags?

While we use user-generated tags as additional content information in this paper, our algorithmic method can in principle be applied to any other type of descriptive item information. We aim at showing that enhancing MF is not only beneficial in terms of model quality, but also with respect to user experience. By employing *TagMF*, users can interactively express their preferences and control the recommendation process in a model-based CF recommender via tags. While ratings stored in an existing user profile or provided during interaction are still taken into account, users can by this means indirectly determine their preferences in the space spanned by the latent factors and interactively adapt the set of recommendations without being required to (re-)rate items. This is possible both in cold-start situations, i.e. for new users entering the system who do not yet have an existing long-term profile, as well as when a profile based on past user feedback is persistently available in the system but the user's needs deviate from long-term interests. Availability of the current user's rating data is not mandatory. Instead, our method requires as input only a conventional dataset of implicit feedback or ratings (of other users) as well as item-related tag relevance

information. From this point of departure, the method allows to derive user-related tag relevance information as well as tag-factor relations. Thus, users themselves do not need to have tagged items before, i.e. we do not require to know a priori how relevant tags are for the current user as we infer this information. Moreover, integrating the easy-to-understand semantics of tags in this novel way allows us to open up the "black box" latent factor models usually constitute for the user. With *TagMF*, we are able to establish a general understanding of the factor space, and to show how users and items are positioned inside it. As a consequence, users can be presented with explicit tag-based explanations of their profile representing preferences they have expressed indirectly with respect to the nontransparent factor space.

To evaluate our method, we first conducted extensive offline experiments comprising an analysis of objective performance and a qualitative inspection of a resulting factor model. Then, in order to validate the application possibilities of *TagMF* and to examine user experience, we implemented a web-based prototype movie RS that uses our method for generating recommendations and for providing users with additional tag-based interaction possibilities. In two quantitative user studies, we compared this interactive system both with a conventional automated recommender that uses MF and with a purely tag-based interactive approach. To the best of our knowledge, our evaluation hence forms the first and most extensive empirical examination of the effects considering additional information has on CF recommenders to date. Among several promising findings, the results indicate that learning an integrated model increases perceived recommendation quality, which previously has only been observed in offline experiments [e.g. 27, 29, 30, 35, 32, 34]. To further analyze aspects related to user experience, we used *Structural Equation Modeling* (SEM) [36]. SEM, a multivariate analysis technique which is still rarely applied in RS research [17, 16], allowed us to investigate the influence applying our method has on the measurement of such aspects and the relationships between them. The analysis yields interesting insights, among others, that users perceive recommendations to be more transparent, and are as a consequence more satisfied with the item finally chosen, when they can additionally interact via tags. In general, the results emphasize the value of considering latent knowledge and (user-generated) content information at the same time—both for improving recommendations and extending interactive control in contemporary RS.

In the following, we first discuss relevant related work. Next, we describe the methodology behind *TagMF* in detail, and elaborate on its application possibilities that allow to implement interactive RS. Afterwards, we present our evaluation, including offline experiments and user studies. Finally, we conclude the paper by discussing the results and providing an outlook on future work.

2. Related Work

Successful examples of commercial recommenders are the systems used by Amazon [11] or Netflix [12], which aim at presenting recommendations that fit well the user's general preferences while reducing interaction effort and cognitive load. However, users might feel too much dominated by the systems, unable to flexibly specify current interests or to obtain, for instance, more diverse and novel recommendations. This is particularly true because users are mostly very limited in their ways to interact with such automated RS or have no control over the recommendation process at all, although this might considerably increase user satisfaction [5, 6, 7, 10]. In contemporary CF recommenders, the only way for users to actively affect the results is usually by providing explicit feedback in form of ratings for single items. While this represents a possibility to exert at least some influence, it does not eliminate the "filter bubble" effect [13] since the user's existing long-term profile is only further refined despite the fact that the search goal may vary depending on the current situation. Moreover, considerable effort is required on part of the user before he or she can obtain adequate suggestions [37, 38]—especially in cold-start situations, i.e. when no historical data is yet available for a new user entering the system or when a user does not want an existing profile to be applied. Apart from that, notably in real-world systems, results are often adapted based on implicit feedback, for example, when users click on interesting items to see more details [3, 2, 10]. This way, user interaction behavior can be modeled more accurately compared to ratings [39, 3], but the process tends to become less transparent and it gets even harder for users to adapt the recommendations towards their situational needs.

2.1. Providing Control and Improving Transparency

In light of these drawbacks, interactive approaches that focus on increasing the level of user control over the recommendation process and improving its transparency have received more and more attention in recent years [40, 41, 10]. For instance, in critique-based RS, users can manipulate the results by critiquing a suggested item with respect to product properties they wish to value higher or lower [21]. In contrast to such attempts, *MovieTuner* [18] does not require previously modeled metadata. Instead, it solely relies on user-generated tags allowing to ask for a movie similar to the currently recommended one—but e.g. less violent and more funny. For tailoring the critiquing process towards the current user, past critiquing sessions can be taken into account [42]. However, long-term profiles as they are customary in CF systems are usually neither considered for adapting the process itself nor do they eventually affect the recommendations.

TasteWeights [22], *SetFusion* [23], *MyMovieMixer* [43] and *uRank* [44] allow to control a RS in a more advanced manner: Users can interactively vary the influence of different social datasources [22], of various algorithms [23], of

certain product facets [43], or of extracted keywords [44] in order to better reflect their current interests. Moreover, these approaches aim at improving system transparency through visualizations that support users in understanding why the items were recommended. Related examples which make even more extensive use of visualization techniques comprise, among others, *MoodPlay* [45] and *Conference Navigator* [46]. A comprehensive overview including attempts to visualize item space and user profiles can be found in [40, 10].

While the most popular type of recommender algorithms is CF [2], many of the attempts proposed to increase interactivity and transparency, including the ones mentioned above, are developed independently of CF: They typically rely on their own concepts to recommend items instead of building on the benefits of established model-based CF techniques that are known for high precision and efficiency [25]. Consequently, even when available, past browsing behavior or previously given ratings cannot be taken into account. Against this background, to our knowledge, no attempts have been made to extend a model-based CF recommender into a fully interactive, user-controlled system.

2.2. Extending Matrix Factorization

Despite the success of MF techniques that learn latent factor models, RS research has been trying to further increase recommendation quality in terms of objective performance metrics [4, 16]. One promising attempt is to complement existing ratings with further data. This additional information may be rather generic, such as implicit user feedback or temporal relations of ratings [24, 25], but often, more specific datasources are taken into account: In [28, 35], predefined content-related metadata about movie genres or recipe ingredients are exploited. Other approaches rely on contextual information, for example, user age or current season [e.g. 27]. Several authors semantically analyze user-written product reviews to first infer hidden topics or opinions about the items, which are subsequently integrated with latent factor models to improve their quality [e.g. 29, 31, 33, 34]. However, only few approaches take immediate advantage of user-generated information such as tags. In these approaches, the underlying models are enhanced with, for instance, specific keywords regarding a movie's mood and plot [30] or generic social tags [32], but are focused on improving offline accuracy rather than user control and system transparency. Accordingly, they have not been evaluated in user studies, leaving the influence on user experience open for investigation. Besides, there exist indeed approaches that rely exclusively on tags for the purpose of generating recommendations, e.g. using graph-based methods [20] or by directly modeling user preferences based on item-tag signals [19, 47]. Yet, they are limited since they cannot benefit from the algorithmic maturity of model-based CF techniques, and thus the availability of existing long-term preference profiles based on implicit or explicit user feedback data. Moreover, apart from e.g. *MovieTuner* [18] or *uRank* [44], these tag-based RS are again not particularly designed for giving users more interactive control.

The range of techniques for considering additional information in CF recommenders is very broad as well. For standard MF [24], which is closely related to *Singular Value Decomposition* (SVD) [48] and thus often referred to as "SVD-like MF", not needing imputation and preventing overfitting by means of regularization [49, 25], a straight-forward way is to add further constraints to the minimization function that is used to learn the parameters when training the latent factor model. This increases precision [24, 32], but after having been learned, the latent factors exhibit no interpretable association with the additional information: The information is calculated into the factor values in a way that the relationship between provided data and latent factors, and consequently items, cannot be made accessible for users anymore. The same applies to approaches that use additional regularization terms [e.g. 29, 30]. In contrast, in [28, 35, 33], the information is explicitly used to establish a content-related association with the factors: By proposing a regression-constrained formulation, factors are considered as functions of content attributes. Further techniques for enriching model-based CF are, among others, extended probabilistic MF [31], deep learning [34, 50], factorization machines [51], or the generalized variant of MF, tensor factorization [27]. All of these attempts have been shown to significantly increase the accuracy objectively measurable in offline experiments. However, to our knowledge, there are currently no empirical user studies available that examine the effects of integrating CF, and in particular latent factor models, with additional information in terms of subjective aspects such as perceived recommendation quality and variety, or user experience in general.

2.3. Exploiting Latent Factor Models

Overall, the usage of latent factor models has rarely been exploited for purposes other than improving effectiveness or performance of RS. Nonetheless, while cold-start situations in CF have mostly been addressed algorithmically [e.g. 52, 53, 54, 55], some exceptions rely on the factor space to interactively elicit initial user preferences, for instance, in a choice-based manner [56, 57]. Likewise, latent factors may contribute to diversify a recommender's output [e.g. 58]. Moreover, notably without the need for explicit content information, they can provide a basis to visualize the item space, e.g. in form of a map [59]. Recently, this metaphor has been extended to a 3D landscape, where the additional dimension represents the current user's interests and allows to interactively express preferences, both with and without an existing profile [60].

In cases where MF has been actually enhanced with additional information as described in Section 2.2, this has primarily served to improve accuracy, not for exposing the additional content information at the user interface. One

of the few exceptions is [61], where user and item characteristics are explained by visualizing the importance of tags according to their correlation with the factors. In [62], a first step towards automatically explaining latent factors in textual form has been taken by associating them with topics inferred from unstructured data. Nevertheless, factors as derived by MF can still be considered overall hard to explain due to their statistical nature, and it seems particularly difficult from a system perspective to relate them to an intelligible meaning [24]. Besides, it can be seen as a more fundamental problem of model-based CF that users typically lack deeper understanding of the underlying mechanisms [5, 7, 15]. For these reasons, latent factor models have yet only rarely been suggested as a means to improve interactive control and transparency in RS.

2.4. Evaluating Recommender Systems

While aspects related to user experience are increasingly considered important for RS research [6, 7, 17], still only few evaluations go beyond measuring performance in retrospective offline experiments [16]. Especially recommenders enhanced with additional information such as tags have not yet been extensively analyzed in empirical user studies. In order to evaluate the user's perception of system and recommendations, the framework proposed in [17] constitutes an important means to explain, among others, how subjective system aspects (e.g. perceived recommendation quality) mediate the impact of objective system aspects (e.g. differences in algorithms) on user experience. Advanced multivariate analysis techniques such as SEM that allow to investigate the underlying relationships are however only rarely used in RS research although they have been considered particularly useful for evaluating user experience [17, 16]. Exceptions have analyzed, for instance, effects of objective system aspects on perception of results [17, 63], influence of choice-based preference elicitation compared to a conventional rating phase [57], how the number of recommended items affects choice difficulty and satisfaction [64], and how diversification based on latent factors may improve these aspects [58]. As already pointed out, it has however not yet been empirically examined how considering additional information in model-based CF actually influences user experience.

2.5. Summary

In summary, it seems promising to extend MF in a way that latent factors can be associated with concepts users understand. Consequently, users could be enabled to interactively control the recommendation process according to their situational needs—both in cold-start situations as well as with an existing preference profile—and be presented with explanations of their formerly opaque representation within the factor model. In this regard, it appears of particular interest to investigate the impact on the subjective assessment of system aspects such as recommendation quality and on user experience in general.

3. Methodology

In this section, we describe *TagMF*, a method to enhance a model-based CF recommender that relies on common user-item interaction data, i.e. implicit feedback or explicit ratings users provided for the items, with additional content information, specifically tags assigned to items by the user community. We show how to learn a model that integrates this item-related tag relevance information in order to subsequently derive corresponding user-tag relevance scores as well as tag-factor relations[1].

In CF, user-item interaction data is usually represented by means of a typically sparse user-item matrix $\mathbf{R} \in \mathbb{R}^{|U|\times|I|}$. By conventional notation, each entry of $\mathbf{R}$ represents a rating r_{ui} given by user $u \in U$ to item $i \in I$, where U is the set of users and I is the set of items [2, 25]. Note that possible values for r_{ui} may differ depending on the application: Typically, the values are numerical ratings (e.g. 1–5), but $\mathbf{R}$ may also contain binary implicit feedback data.

Standard SVD-like MF (see Section 2.2) reduces the dimensionality of $\mathbf{R}$ by learning a latent factor model which then serves to generate recommendations [24, 25]. This model approximates $\mathbf{R}$ through two low-rank matrices, $\mathbf{P} \in \mathbb{R}^{|U|\times|F|}$ and $\mathbf{Q} \in \mathbb{R}^{|I|\times|F|}$, where F is a set of latent factors[2]. The user-factor matrix $\mathbf{P}$ and the item-factor matrix $\mathbf{Q}$ can be trained using optimization algorithms such as Stochastic Gradient Descent or Alternating Least Squares, which are able to efficiently handle sparse matrices [24, 25]. A user's u (calculated) interest in a particular factor f is then numerically expressed by entry p_{uf} of $\mathbf{P}$ while entry q_{if} of $\mathbf{Q}$ describes the extent to which item i possesses this factor. Consequently, with users and items being mapped into the same factor space [24], the inner product of a user-factor vector p_u and an item-factor vector q_i captures the interaction between user and item, and thus allows to predict the rating $\hat{r}_{ui}$ of user u for item i. Overall, this results in:

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q}^{\mathrm{T}} \tag{1}$$

As our method is in principle independent of algorithmic details, we omit elaborating on e.g. regularization terms and refer to the literature [e.g. 24, 25] for a general and more extensive introduction to MF.

3.1. Integrating Item-Related Tag Information

Since a latent factor model derived as described above cannot be directly integrated with additional information, we need to add further constraints. Initially following the approach proposed in [28] (see Section 2.2), we complement a SVD-like MF algorithm by extending $\mathbf{Q}$ with item-related information. For this, we use tag relevance scores

[1] The basic principle of our method was introduced in the poster publication of [65]. Now, we describe the method in more detail and subsequently discuss the possibilities to actually apply *TagMF* in common model-based CF systems.

[2] The number of factors (typically 10 to 100) has to be specified before the actual factorization.

for items relying on a set of tags T, and define ${}^{\mathbf{i}}\mathbf{A} \in \mathbb{R}^{|I|\times|T|}$ as a matrix representing how strongly items relate to tags: Each entry a_{it} of ${}^{\mathbf{i}}\mathbf{A}$ describes on a continuous scale from 0 (not relevant) to 1 (very relevant) the degree to which a tag t is relevant for an item i. However, we additionally extend $\mathbf{P}$ and define ${}^{\mathbf{u}}\mathbf{A} \in \mathbb{R}^{|U|\times|T|}$ to also represent user-tag relations, i.e. tag relevance scores for users. From that, we redefine the original MF model given in (1) as follows:

$$\mathbf{R} \approx \mathbf{P}\mathbf{Q}^{\mathrm{T}} = {}^{\mathbf{u}}\mathbf{A}\,{}^{\mathbf{u}}\mathbf{\Theta}({}^{\mathbf{i}}\mathbf{A}\,{}^{\mathbf{i}}\mathbf{\Theta})^{\mathrm{T}}, \tag{2}$$

where ${}^{\mathbf{u}}\mathbf{\Theta} \in \mathbb{R}^{|T|\times|F|}$ associates tags with factors as seen from user side and ${}^{\mathbf{i}}\mathbf{\Theta} \in \mathbb{R}^{|T|\times|F|}$ is the equivalent for items. In fact, this represents a regression-constrained formulation of the MF problem, where each factor is a function of the content attributes.

Additional content information may only be available either for users or for items. In [28], for instance, content-related metadata explicitly defined for the items has been taken into account (see Section 2.2). Since we aim at enhancing MF with tags users provided for the items, we assume that item-related tag relevance information is known a priori, and the corresponding matrix ${}^{\mathbf{i}}\mathbf{A}$ has been determined separately with a suitable method. In principle, this relationship between items and additional information can be quantified using any type of attribute that relates to both user information space and item information space in a meaningful way. The only requirement is that a numerical representation can be derived so that the entries of ${}^{\mathbf{i}}\mathbf{A}$ hold the respective relevance scores for items on a continuous scale. Information on which specific users applied which tags is however not required a priori: In contrast to matrix ${}^{\mathbf{i}}\mathbf{A}$, we consider the corresponding matrix for users, ${}^{\mathbf{u}}\mathbf{A}$, to be unknown. Consequently, we treat the whole term ${}^{\mathbf{u}}\mathbf{A}\,{}^{\mathbf{u}}\mathbf{\Theta}$ implicitly at this step by just learning the user-factor matrix $\mathbf{P}$ as known from standard MF. With this constrained equation, we can now formulate the following minimization problem as done in [28]:

$$\min_{\mathbf{P},\,{}^{\mathbf{i}}\mathbf{\Theta}} \sum_{(u,i)\in K} (r_{ui} - p_u^{\mathrm{T}}\,{}^{\mathbf{i}}\mathbf{\Theta}^{\mathrm{T}} a_i)^2 + \lambda\big(\sum_{u\in U}\|p_u\|^2 + \|{}^{\mathbf{i}}\mathbf{\Theta}\|^2\big), \tag{3}$$

with λ controlling the extent of regularization and K being the set of all user-item tuples for which user feedback (e.g. ratings) exists. We then apply a gradient descent algorithm with learning rate μ to minimize the error:

$$\begin{aligned} p_u &\leftarrow p_u + \mu\Big(\sum_{i\in K_u}(r_{ui} - p_u^{\mathrm{T}}\,{}^{\mathbf{i}}\mathbf{\Theta}^{\mathrm{T}} a_i)\,{}^{\mathbf{i}}\mathbf{\Theta}^{\mathrm{T}} a_i - \lambda p_u\Big) \\ {}^{\mathbf{i}}\mathbf{\Theta} &\leftarrow {}^{\mathbf{i}}\mathbf{\Theta} + \mu\Big(\sum_{(u,i)\in K}(r_{ui} - p_u^{\mathrm{T}}\,{}^{\mathbf{i}}\mathbf{\Theta}^{\mathrm{T}} a_i) a_i p_u^{\mathrm{T}} - \lambda\,{}^{\mathbf{i}}\mathbf{\Theta}\Big) \end{aligned} \tag{4}$$

3.2. Deriving Tag-Factor Relations for Users

At this point, we have transferred the abstract factor semantics into a comprehensible information space utilizing a regression-constrained approach on the item side. Although we considered tag relevance scores to be known only for items, we can now establish a relationship between users and tags, enabling us later to let users specify their interests via tags and to explain their profile to them.

For this purpose, we apply the learned relationship between tags and latent factors to the user side. This is possible as the way a MF model is learned (see above) ensures per definition that both users and items are mapped into a joint factor space [24]. Thus, each factor $f \in F$ reflects a certain characteristic that has the same (hidden) semantic meaning for both users and items [24, 26]. The regression coefficients hence describe tag-factor relations in general, for users as well as for items. Accordingly, the implicitly assumed ${}^{\mathbf{u}}\mathbf{\Theta}$ is equivalent to ${}^{\mathbf{i}}\mathbf{\Theta}$, such that:

$${}^{\mathbf{u}}\mathbf{\Theta} = {}^{\mathbf{i}}\mathbf{\Theta} =: \mathbf{\Theta} \tag{5}$$

As a consequence, ${}^{\mathbf{u}}\mathbf{A}$ is now the only unknown left. Based on our problem formulation, its row vectors a_u should hold the equivalents of the item-related tag relevance scores from ${}^{\mathbf{i}}\mathbf{A}$ with respect to users. In accordance with (2), we thus solve for ${}^{\mathbf{u}}\mathbf{A}$:

$$\begin{aligned} \mathbf{P} &= {}^{\mathbf{u}}\mathbf{A}\mathbf{\Theta} && \Leftrightarrow \\ \mathbf{P} &= {}^{\mathbf{u}}\mathbf{A}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} && \Leftrightarrow \\ {}^{\mathbf{u}}\mathbf{A} &= \mathbf{P}\mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\mathrm{T}} && \Leftrightarrow \\ {}^{\mathbf{u}}\mathbf{A} &= \mathbf{P}\mathbf{\Theta}^{+} \end{aligned} \tag{6}$$

Since $\mathbf{\Theta}$ is generally not a square matrix, we have to calculate its pseudoinverse $\mathbf{\Theta}^{+}$ (i.e. the Moore-Penrose generalization of the inverse matrix [66, 67]) first by applying SVD [48], yielding $\mathbf{U} \in \mathbb{R}^{|T|\times|T|}$, $\mathbf{\Sigma} \in \mathbb{R}^{|T|\times|F|}$ and $\mathbf{V} \in \mathbb{R}^{|F|\times|F|}$. Consequently, $\mathbf{\Theta}^{+}$ is defined as $\mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^{\mathrm{T}}$.

The general interest of user u with respect to all tags provided by the user community is now expressed by vector a_u of ${}^{\mathbf{u}}\mathbf{A}$, which is easy to understand and basically the calculated counterpart of the given item-tag relevance scores introduced in Section 3.1.

Finally, since $\mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}}$ holding the general tag-factor relations in (2) is a square diagonalizable matrix, we can represent it in terms of eigenvalues and eigenvectors using eigendecomposition:

$$\begin{aligned} \mathbf{R} &\approx {}^{\mathbf{u}}\mathbf{A}\mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}}\,{}^{\mathbf{i}}\mathbf{A}^{\mathrm{T}} \\ &\approx {}^{\mathbf{u}}\mathbf{A}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\,{}^{\mathbf{i}}\mathbf{A}^{\mathrm{T}} \\ &\approx {}^{\mathbf{u}}\mathbf{A}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}\,{}^{\mathbf{i}}\mathbf{A}^{\mathrm{T}} \end{aligned} \tag{7}$$

The diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues of $\mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}}$ in non-increasing order. The eigenvectors in $\mathbf{U}$ hold the importance of every tag with respect to a certain direction. Since $\mathbf{\Theta}\mathbf{\Theta}^{\mathrm{T}}$ is symmetric, eigenvectors are chosen orthogonal to each other. Latent factors are thus incorporated into the tag information space by stretching it along the eigenvector directions according to the magnitude of the corresponding eigenvalues.

4. Application Possibilities

In this section, we describe several ways *TagMF* can be applied so that users may take benefit of tags in RS that rely on common model-based CF. The integrated model of latent factors and additional content information derived using our proposed method gives us the opportunity to access the previously abstract user-factor and item-factor vectors in a much more comprehensible manner: User profiles and item descriptions now comprise information related to both latent factors and user-generated tags. Thus, as the tag concept is easily understood by users, we can exploit the enriched vectors for several purposes: Among others, users may actively adjust their own user vector, i.e. indirectly determine their position in the latent factor space, in an interactive manner by means of tags according to their current situation. More concretely, in relation to the research questions formulated in Section 1, we enable users to ...

- select a small number of tags to express preferences at cold-start instead of rating items up front (RQ1a),
- weight tags to manipulate recommendations generated based on their existing user profile (RQ1b),
- critique a recommended item to receive suggestions that also take their profile into account (RQ1c),
- examine their preference profile by means of tag-based explanations (RQ1d).

In the following, we describe in detail how *TagMF* can be applied to realize interactive RS that support users in the different cases.

4.1. Eliciting Preferences at Cold-Start

In cold-start situations, users typically have to rate a certain number of items before CF recommenders can reliably predict their interests [38, 55] (see Section 2). When employing *TagMF*, new users can, in contrast to a conventional preference elicitation phase, be asked to select a (small) number of preferred tags to establish a user profile.

For this, we initialize a new user-tag vector a_u for a user u entering the system as follows:

$$a_{ut} = \begin{cases} 1 & \text{if tag } t \text{ has been selected by user } u \\ 0 & \text{else} \end{cases} \tag{8}$$

By multiplying this vector a_u with $\mathbf{U}\mathbf{\Lambda}^{1/2}$ (see (7)) holding the tag-factor relations, we obtain a regular user-factor vector. Now, to generate recommendations, this vector $a_u\mathbf{U}\mathbf{\Lambda}^{1/2}$ can be used the same way as if the vector p_u representing the user profile in standard MF had been derived exclusively based on ratings. This means we calculate its inner product with the item-factor vectors as shown in the introductory description in Section 3 (see also [24, 25]).

4.2. Manipulating Recommendations

For a user u with an existing preference profile based on explicit ratings or implicit behavioral data, i.e. a vector p_u is already available, usually the only means to influence the recommendations in model-based CF systems is to (re-)rate single items (see Section 2). However, when a_u is derived by *TagMF* in the learning phase as described in Section 3.2, the user can additionally manipulate the entire result set in an interactive manner by means of tags provided by the community of users. This may support users in obtaining alternative suggestions, for instance, in case their long-term profile differs from actual interests or the recommendation list lacks diversity and novelty.

To this end, we define a weight vector $w_u \in [0,1]^{|T|}$ that is supposed to capture user feedback in form of weights for tags, where 0 means no and 1 maximal interest of user u in tag t. For instance, a user who in the current situation is interested in action-packed movies that moreover contain a little more black humor than the ones usually recommended to him or her, may set the weights of the tags "action" and "black humor" to 1 and 0.5, respectively. This vector w_u can be then added to a_u in order to calculate recommendations based upon this update to the existing user profile. Consequently, we extend the original formulation (see again Section 3 as well as [24, 25]) as follows:

$$\tilde{r}_{ui} = (a_u + \pi w_u)\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} a_i, \tag{9}$$

with $\pi \in \mathbb{R}$ representing the degree to which the weight information is considered: We adaptively set π to 0 when the current user has not applied any weights, otherwise to $\frac{\|a_u\|}{n} \cdot \|w_u\|$, with $n \leq |T|$ being the number of tags already weighted by the user. Thus, when he or she sets all n weights to the maximum value, they have the same influence as a_u itself, i.e. both vectors are of equal length.

Provided users have any means to manipulate the values of w_u, e.g. sliders or spinners, the set of recommendations initially generated based on their long-term profile can now be continuously adapted in realtime, allowing them to interactively explore the effects of their preference settings, and, among others, to escape a potential "filter bubble". Thereby, we in principle do not longer predict actual ratings: Only at the beginning, when all values of w_u are set to 0, $\tilde{r}_{ui}$ effectively approximates r_{ui}. Instead, we combine the user's general preference structure with the operationalization of his or her current interests or situational (e.g. mood- or activity-based) needs w_u, that he or she has expressed with respect to the tags by interacting with the system.

4.3. Critiquing a Recommended Item

Employing *TagMF* makes it further possible for users to interact with model-based CF systems in a more discrete fashion, resembling the well-known critiquing approach [21] (see Section 2.1). As in *MovieTuner* [18], an interactive variant based on user-generated tags implemented as part of the MovieLens[3] platform, we are able

[3] `https://movielens.org/`

to let users request items that are overall similar to a currently recommended item i, but represent some selected dimensions less, equally or more strongly. This way, specific context-dependent or situational aspects of the search and decision process can be taken into account. For instance, in case the movie "Apocalypse Now" is shown, a user might apply the tag-based critique "less dark", leading to "Saving Private Ryan" being suggested.

However, since our method builds on MF, we can additionally exploit the current user's long-term profile. As a consequence, results presented after critiquing are not only related to the critiqued item (generally similar, but different with respect to applied critiques), but take to some degree this user's general interests inferred from past user-item interaction data into account as it is customary for CF recommenders. Thus, considering the example from above, it might be that a user who tends to enjoy comedy more than other genres is presented with e.g. the movie "M*A*S*H" instead of "Saving Private Ryan" as a new recommendation. Moreover, the latent information available when using *TagMF* may influence the critiquing process in a way that resulting recommendations also reflect more subtle item characteristics that cannot be taken into account solely relying on explicit tag data.

Eventually, on condition that meaningful tags are somehow selected and presented as critique dimensions, it is necessary to reflect the critiques a user u has applied with respect to these tags to the currently recommended item i. For the implementation of this interaction mechanism, we combine the item-tag vector a_i with the user-tag vector a_u to a new vector a_c by performing the following steps[4]:

1. We scale a_i to the length of a_u, yielding a'_i. This ensures that in the end, we can still use a_c on the user side for generating recommendations.
2. Assuming that u likes the current suggestion due to very specific characteristics of i, we keep only values of a'_i that are two standard deviations above the mean of a'_i. All other entries are set to 0. Thereby, we avoid too homogeneous entries in a_c as it might be the case when just directly averaging all values of a'_i and a_u to combine them, which would lead to results neither related to i's characteristics nor u's profile.
3. We use a weighted average to combine a'_i with a_u, integrating a'_i with higher weight (here 60 %) in order to more strongly reflect i's similarity to the items in the new result set. As the critiquing process continues, the weights may be dynamically adjusted.

Now, to generate recommendations, the resulting vector a_c can be used the same way as a_u before (see previous subsections). These recommendations are simultaneously geared towards i's characteristics as well as u's general interests regarding the tags. To also fulfill the user's critique he or she has interactively applied, we employ the *linear-sat* variant of the critique distance (i.e. the difference of i along the selected critique dimensions to the other items) as proposed in [18].

[4]We decided for the reported configuration due to pretests.

4.4. Explaining a User Profile

In systems relying on MF, users typically express their preferences indirectly with respect to the nontransparent latent factor space, e.g. through ratings for single items. The result are abstract user-factor vectors, making it difficult to explain a user's profile. This can also be considered a common and more general issue in model-based CF. Our method, in contrast, allows to provide users with explicit tag-based explanations of the typically opaque representation of their long-term preferences within the model: As a consequence of taking additional content information into account, we can automatically determine those tags that are most important to an individual user—even if he or she never tagged any items.

For this purpose, we exploit that with *TagMF*, user-factor vectors are related to both latent knowledge and user-generated content, and thus become much more meaningful. Concretely, we utilize the matrix ${}^{\mathbf{u}}\mathbf{A}$ holding the user-tag relations in order to explain the user representation as learned from historical user-item interaction data in textual form. When ${}^{\mathbf{u}}\mathbf{A}$ is derived according to our method, this is independent of the tags a specific user actually has assigned: As described in Section 3.2, we derive tag-factor relations for all users by first learning the relationship between tags and latent factors, and then applying it to the user side. Hence, we can identify the most important tags for each user, even in the common case where he or she has not provided any tags but only conventional feedback (e.g. ratings). Thus, for the current user u, we select the n tags scoring highest in the corresponding user-tag vector a_u, and present them as a description of his or her long-term interest profile.

5. Evaluation

In order to answer the research questions posed in Section 1, we extensively evaluated *TagMF* both in offline experiments and empirical user studies.

First, to provide a basis for addressing RQ1, we performed offline experiments comprising an analysis of objective performance as well as a qualitative inspection of a resulting latent factor model. These experiments were supposed to show validity and general effectiveness of our method for enhancing model-based CF with additional content information.

Next, to analyze our method's actual impact on users and to investigate the extended interaction possibilities provided, we conducted two empirical user studies: In the first study, we focused on the usage of *TagMF* for eliciting preferences in cold-start situations (RQ1a) and interactively manipulating recommendations that result from an existing user profile (RQ1b). In the second study, we investigated how *TagMF* can be applied to integrate model-based CF with critiquing, taking the recommended item

as well as the current user's long-term interests into account (RQ1c). For these quantitative studies, we built an interactive web-based prototype movie RS that implements *TagMF*. To specifically examine the influence on subjective system aspects and user experience, we then performed a comparison with an automated recommender based on standard MF using ratings (RQ2a) in the first study, and with a tag-based interactive approach similar to *MovieTuner* (RQ2b) in the second one.

In the following, we describe all parts of this three-fold evaluation, concluding each with a detailed discussion that addresses the respective research questions.

5.1. Offline Experiments

Earlier experiments by others (see Section 2.2 and [e.g. 27, 29, 30, 35, 32, 34]) suggested that considering additional information improves accuracy of model-based CF recommendations as measured by common offline evaluation metrics. To confirm these findings and to validate our method's effectiveness, we as well analyzed the objective performance[5].

Moreover, while latent factors are generally considered to represent real-world characteristics [24, 26], we conducted a qualitative inspection of a factor model derived by means of *TagMF* to investigate whether automatically learning tag-factor relations according to our method actually leads to comprehensible and meaningful results.

5.1.1. Setup

In order to perform the experiments we used a Stochastic Gradient Descent MF algorithm[6] based on [68] as a baseline. We extended this implementation of a common SVD-like MF algorithm according to our method as described in Section 3. As datasource for items as well as associated ratings and user-generated tags, we used the well-known MovieLens 20M dataset for ratings and the MovieLens Tag Genome dataset for item-tag relevance scores[7]. We then created an intersection of these datasets reducing them to items included in both, leaving us with 10 370 movies, 19 800 443 ratings and 11 697 360 tag relevance scores.

To run the performance analysis, i.e. to objectively compare standard SVD-like MF with *TagMF* in terms of recommendation accuracy, we used the RiVal benchmarking toolkit[8] introduced in [69]. With this toolkit, we computed *Root Mean Square Error* (RMSE) [4] and *Normalized Discounted Cumulative Gain* (NDCG) [4], a popular ranking metric from Information Retrieval.

5.1.2. Analysis of Objective Performance

First, we examined the influence of different basic configurations on objective recommendation accuracy using 10 % subsamples of users and 5-fold cross validation. We trained the standard MF and the *TagMF* models with 20 factors. For *TagMF*, we considered a limited number of the 50 most popular user-generated tags from the underlying dataset as additional training data. Figure 1 shows the experimental results for a comparison of standard MF and *TagMF* in terms of RMSE and NDCG@10 which we calculated as described above, varying the number of iterations and the regularization parameter λ when training the respective model.

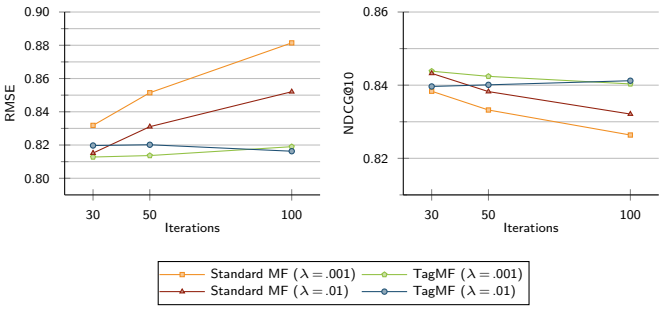

Figure 1: Comparison of standard MF and *TagMF* in terms of RMSE and NDCG@10 for different number of iterations and settings for λ.

Looking at these results, it seems that enhancing MF with additional information according to our method is beneficial. *TagMF* yields overall superior results both in terms of RMSE and NGCG@10. Furthermore, the results obtained with *TagMF* are rather stable. In contrast, iterating more often over the training data leads to decreased performance for standard MF.

Second, with 1 % subsamples of users, we performed another comparison of standard MF with *TagMF*, now varying the number of latent factors and the number of tags additionally considered in *TagMF*. Following further pretests, we used 30 iterations and set $\lambda = .03$. Then, we again performed 5-fold cross validation, yielding the RMSE and NDCG@10 results reported in Figure 2.

Overall, it again becomes apparent that considering additional information improves objective accuracy of MF. When using 50 tags or more, RMSE is lower for *TagMF* independent of the number of latent factors. NDCG@10 shows similar behavior, yielding equally promising results.

5.1.3. Qualitative Inspection

Enhancing a model-based CF recommender with additional content information according to our method may also help to gain a better understanding of the latent factor space[9]. Applying eigendecomposition as described in

[5] First results of offline experiments have been shown in the poster publication of [65]. Now, we present additional and more extensive experiments, among others with a newer and larger dataset.

[6] *ParallelSGDFactorizer* from the Apache Mahout recommender library (`http://mahout.apache.org/`).

[7] The MovieLens 20M dataset contains about 20 million ratings from more than 138 000 users for over 27 000 movies; The MovieLens Tag Genome dataset contains item-tag relevance scores for over 10 000 movies and 1 100 user-generated tags (`https://grouplens.org/datasets/`).

[8] `http://rival.recommenders.net/`

[9] We have briefly discussed this in the poster publication of [26].

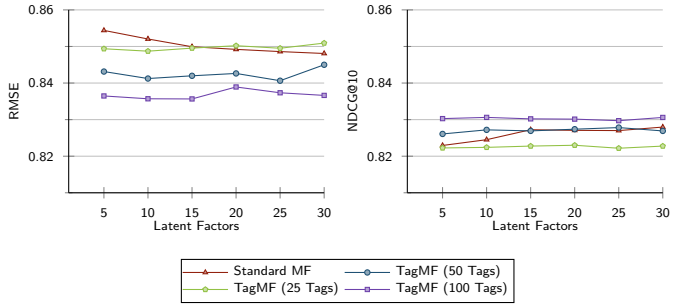

Figure 2: Comparison of standard MF and *TagMF* in terms of RMSE and NDCG@10 for different number of latent factors and tags.

Section 3.2 yields information on the importance of each dimension of the factor space and its relationship to the tags. Consequently, by examining the most positively and negatively related tags, respectively, we can obtain a more general understanding of what is expressed by factors derived automatically by means of MF.

Table 1 illustrates an example for relationships learned between factors and tags, resulting from a *TagMF* model trained on the MovieLens 20M dataset with 20 factors and 20 user-generated tags: Positive and negative values describe strength and direction of these relations. They express how strongly certain characteristics are represented within the respective factors, thus denoting their individual meaning. Apparently, the underlying semantics can be easily interpreted: For instance, while both factor 4 and 5 express characteristics related to "fantasy", factor 4 has a very negative and factor 5 a very positive relation to the tag "action". Accordingly, these two factors correspond to very different kinds of fantasy movies. This observation can be underpinned by extracting representative items for the respective factors, i.e. which have highest values in the item-factor matrix ${}^{\mathbf{i}}\mathbf{A}\mathbf{U}\mathbf{\Lambda}^{1/2}$ (see Section 3.2). Here, for example, "Wizard of Oz" (factor 4) and "Star Wars: Episode IV – A New Hope" (factor 5) are clearly in line with the observed semantics.

5.1.4. Discussion

The offline evaluation generally shows that enhancing MF with additional information seems indeed beneficial in terms of objective recommendation quality. This is consistent with earlier retrospective offline experiments (see Section 2.2), validating the work of others [e.g. 27, 29, 30, 35, 32, 34] and providing a basis to further investigate RQ1.

With the rather limited subsamples of rating data used in our analysis, the decreasing accuracy of standard MF in Figure 1 compared to the largely stable results of *TagMF* is likely attributed to overfitting: Additional tag-based information appears to contribute more to control overfitting than increasing λ for standard MF. Also, as can be seen in Figure 2, the number of latent factors clearly has an influence on the performance of standard MF: The results improve with more factors and become stable only with 15 to 20 factors, while this parameter does not seem to affect *TagMF* to a large degree. With the amount of training data used, the number of tags incorporated according to our method seems to be the predominant factor for model quality. Nevertheless, with few tags (25 and 50), RMSE for *TagMF* goes up slightly when increasing the number of factors. Apparently, the variance in the factors cannot be covered sufficiently by the tags when there are fewer tags than factors. Thus, more factors appear to require considerably more tags to ensure consistently high model quality. Accordingly, in the example in Table 1, each factor is strongly related to multiple tags. However, parameter tuning in general, and e.g. determining an optimal ratio of factors to tags, is subject to future work. Overall, while observed differences between standard MF and *TagMF* are rather small independent of the number of tags taken into account (see again Figure 2), one can expect them to significantly increase when using a larger set of ratings. Then, as the data-generating function gets more complex, including more factor dimensions can be assumed to gain impact.

The qualitative inspection of the integrated *TagMF* model suggests that additional information may actually contribute to opening up the "black box" such models usually constitute for users in CF systems. The derived relations seem useful for the purpose of explaining latent factors through easily comprehensible user-generated tags. Moreover, the regression-constrained formulation (see Section 3.1) allows to gain insights on how users and items are positioned inside the latent space.

As shown in Table 1, we found items to be representative for the different dimensions and their relationship to the tags. With *TagMF*, latent factors are related to the tag information space by eigenvectors and eigenvalues (see Section 3.2), making it possible to translate positions for users the same way as for items. Thus, the method we use to derive user-tag relations (see again Section 3.2) ensures that users are assigned to equally meaningful positions. Accordingly, we proposed in Section 4.4 how to exploit these user-tag relations to select tags that explain a user's long-term preference profile, thereby addressing the corresponding research question (RQ1d). While our approach consequently appears to be indeed a promising means to present users with explicit tag-based descriptions of their—in model-based CF typically nontransparent—preference profile, further validating this application possibility seems necessary.

5.2. Empirical User Study I

We performed the first user study to examine the influence considering additional information has on users in model-based CF systems, and to evaluate the interactive features that become possible by using *TagMF* in comparison to a conventional recommendation process[10].

[10]This study has in large parts been presented in [70]. Now, we present more results and additional insights.

Table 1: Example for automatically learned relationships between latent factors (rows) and user-generated tags (columns): The five most important factors are shown together with positively (green) and negatively (red) related tags, as indicated by $\mathbf{U}$. The factor importance (in brackets in the left-most column) is equal to the values in $\mathbf{\Lambda}^{1/2}$. Representatives for each factor are automatically determined by extracting the movies (with at least 10 000 ratings) that score highest for the respective factor in the actual item-factor matrix ${}^{\mathrm{I}}\mathbf{A}\mathbf{U}\mathbf{\Lambda}^{1/2}$.

Factor	action	atmospheric	based on a book	classic	comedy	dark comedy	disturbing	dystopia	fantasy	funny	psychology	quirky	romance	sci-fi	surreal	time travel	thought-provoking	twist ending	violence	visually appealing	Automatically extracted representatives
1 (1.66)	0.25	0.38	-0.14	0.47	-0.20	0.16	0.14	0.04	-0.27	-0.15	-0.09	0.17	-0.26	-0.03	0.15	-0.19	0.06	-0.36	0.11	0.24	The Shining, Taxi Driver, A Clockwork Orange
2 (1.51)	-0.11	-0.12	0.02	-0.34	0.12	0.21	0.26	0.12	0.27	-0.13	-0.02	0.14	-0.30	0.22	0.36	-0.51	-0.06	0.09	0.14	-0.21	Natural Born Killers, Brazil, Beetlejuice
3 (1.30)	0.10	0.11	-0.13	-0.63	-0.10	0.11	-0.06	0.07	-0.16	0.06	-0.07	0.21	-0.03	-0.16	0.18	0.17	-0.11	-0.05	-0.07	0.59	Amélie, Sin City, Magnolia
4 (1.21)	-0.39	-0.06	0.24	0.12	-0.16	0.00	0.03	-0.12	0.50	-0.22	0.05	0.14	0.29	-0.17	0.17	0.02	-0.08	-0.04	-0.48	0.19	Wizard of Oz, Willy Wonka & the Chocolate Factory, The NeverEnding Story
5 (1.17)	0.44	0.17	0.01	0.13	-0.11	-0.29	-0.16	0.01	0.44	0.10	-0.12	-0.27	-0.42	0.15	0.02	-0.16	0.01	-0.05	-0.20	0.28	Star Wars: Episode IV – A New Hope, Hobbit: An Unexpected Journey, Thor: The Dark World

5.2.1. Goals

First, we laid our focus on validating application possibilities of *TagMF*: For examining the value of user-generated tags as a means to elicit preferences at cold-start (RQ1a) and to interactively manipulate recommendations based on an existing user profile (RQ1b), we implemented them according to Section 4.1 and 4.2, respectively, in our a web-based prototype movie RS. Next, since we were interested in comparing the impact of additional information on subjective system aspects and resulting user experience to an automated rating-based CF recommender as it is common today (RQ2a), we formulated the following hypotheses contrasting this baseline and *TagMF*:

- **H1:** *TagMF* improves perceived quality of recommendations.
- **H2:** *TagMF* improves satisfaction with the item chosen from the recommendations.
- **H3:** *TagMF* decreases difficulty to choose an item.
- **H4:** *TagMF* has no negative impact on perceived interaction effort.
- **H5:** *TagMF* improves transparency, especially in cold-start situations.

5.2.2. Method

The study was designed as an experiment under controlled conditions. We recruited 46 participants (33 female) with an average age of 22.89 ($SD = 6.88$), most of them students (85 %). To interact with the prototype RS and to answer questionnaire items, participants used a common web browser at a desktop PC with a 24" LCD (1920×1200 px resolution). In the following, we describe the prototype system, the procedure, and the questionnaire we used, in more detail.

Prototype. Figure 3 shows the web-based prototype movie RS we implemented for the first user study. We set up two variants: One with a standard SVD-like MF algorithm [24], allowing users only to rate items. The interface resembled a typical automated recommender based on ratings, with no interface elements related to tags present. This variant served as a baseline to test our hypotheses. The other variant was implemented based on *TagMF*. In order to validate the application possibilities, we extended this variant in comparison to contemporary model-based CF systems with several tag-based interaction mechanisms, as described in Section 4.1 and 4.2.

Concretely, the interface of the prototype is structured as follows: At the top (a), an area is shown where—in the *TagMF* variant—users can place tags and subsequently adjust their weight by means of sliders attached to them, this way manipulating the values of w_u (see Section 4.2). Note that in our prototype, it is not possible for users to create tags themselves, but only to use tags from the underlying dataset of tags provided by other users (see below). As *TagMF* can easily be applied with any set of tags, including tags generated by users of the respective system, this would indeed be different in a real-world scenario. Below (b), an input field allows to manually search for tags other users have applied, supported by autocompletion. These tags may be chosen to be weighted, i.e. to be placed in the area at the top together with a slider. In addition, the system initially suggests the 7 most popular tags, i.e. that have been assigned most often by users. As soon as the current user weights some tags, tags similar in terms of item-tag relevance data are suggested. The dialog in the top-right corner (c) presents users with a tag cloud describing their existing preference profile by means of tags chosen as described in Section 4.4.

Beneath, independent of the variant, the top-10 recommended items (d) are displayed together with movie poster and metadata. To further refine their profile, users may rate recommended movies and explicitly search further titles in order to rate them as well. In the *TagMF* variant, alongside each recommendation, the 3 most relevant tags for the respective movie are additionally shown (which may also be chosen to be weighted). Each manip-

ulation updates the result set immediately, thus providing users with direct and meaningful feedback regarding the effects of their preference settings on the recommender.

For calculating recommendations based on ratings, we used the same baseline algorithm as in the offline evaluation (Section 5.1.1). For the *TagMF* variant, we extended this algorithm according to Section 3, and implemented the interactive features as described in Section 4. Pretests similar to the offline experiments presented in Section 5.1, but based on the MovieLens 10M dataset[11], suggested to use 20 factors, 40 iterations, $\lambda = .001$, and to consider the 25 most popular user-generated tags from the underlying dataset as additional training data. We used the MovieLens 10M dataset for ratings and the MovieLens Tag Genome dataset for associated tags[12]. We created an intersection of these datasets reducing them to those movies included in both, leaving us with 8 429 items, 9 964 745 ratings and 9 507 912 tag relevance scores. For the purpose of the study, we used scores precomputed as described in [71] based on user-generated tags from the underlying dataset. In a real-world scenario, one would indeed calculate the scores based on tags provided by the user community of the respective system, and then apply *TagMF* accordingly.

Questionnaire and Log Data. The questionnaire participants were required to fill in was primarily based on the pragmatic evaluation procedure for RS described in [72], containing items related to subjective system aspects and user experience. This evaluation framework (see Section 2.4) is based on [17], but is reduced to stable operationalizations of the subjective constructs and appears (after repeatedly being validated) to measure user experience in RS reasonably well with a limited number of questionnaire items [72]. Concretely, we assessed *Perceived Recommendation Quality*, *Choice Satisfaction*, *Choice Difficulty* and *Effort* by means of items from this framework. We used an additional item from [73] to assess recommendation *Transparency*.

We generated items ourselves to explicitly ask which of the two variants of the prototype RS participants prefer in general, and to let them rate the suitability for different situations of use (with or without search goal). To specifically analyze the usability of the additional interaction mechanisms, we applied the *System Usability Scale* (SUS [74]) and the *User Experience Questionnaire* (UEQ [75]) for the *TagMF* variant. In addition, we used again items from [73] to assess interface adequacy. Besides, we gathered data about demographics, interest in movies, and familiarity with the movie domain. Apart from UEQ (7-point bipolar scale), all items were assessed on a positive 5-point Likert-scale (1–5). We also collected qualitative feedback: An open-ended question asked participants to report suggestions and complaints. We logged user interaction behavior and measured task times.

Procedure. First, each participant was asked to complete two preliminary tasks in counter-balanced order that served to elicit an initial set of preferences both in form of numerical ratings (like in other CF systems) and preferred tags:

a) Participants had to rate 10 out of the 30 most popular movies, which is a common value for a number of ratings that already leads to appropriate results [38]. We used these ratings for online-updating the factor vectors as proposed in [49]. Items were shown in random order and could be skipped when unknown.
b) Participants had to select 3 tags[13] they liked out of the 20 most popular ones from the dataset (shown in random order), which we used to initialize a meaningful user-tag vector a_u as described in Section 4.1.

Next, based on the two system variants implementing a standard MF algorithm and *TagMF*, respectively, we assigned participants in counter-balanced order to three different conditions in a within-subject design:

Standard MF: Standard SVD-like MF with initial recommendations based on the 10 ratings. The only interaction possible was to rate more items.

***TagMF*-Ratings:** Tag-enhanced MF with initial recommendations based on the 10 ratings. Participants could again rate more items, but in addition weight tags in an interactive manner.

***TagMF*-Tags:** Tag-enhanced MF with initial recommendations based on the 3 selected tags. Interaction mechanisms were equivalent to the previous condition.

In each condition, participants were initially presented with the top-6 recommendations generated by means of the respective algorithm. First, they were asked to choose one movie from these suggestions they would actually like to watch. Second, they rated their satisfaction with each of the movies on a 5-point Likert-scale (1–5). Third, they filled in the questionnaire described above regarding their subjective assessment of system and recommendations.

Next, in the interaction phase, participants were presented with the interface of the prototype variant that corresponds to the respective condition, showing the top-10 recommended movies (see Figure 3). Their task was to interact with the system using the provided means in order to further refine the recommendations and to receive a result set that better matched their personal interests. Eventually, participants finished the interaction phase at their own discretion.

[11] At the time we conducted the first user study, not all data was yet released for the MovieLens 20M dataset.

[12] The MovieLens 10M dataset contains about 10 million ratings from more than 70 000 users for over 10 000 movies; The MovieLens Tag Genome dataset contains item-tag relevance scores for over 10 000 movies and 1 100 user-generated tags (`http://grouplens.org/datasets/`).

[13] For the number of tags to be selected, we analyzed the general interest of all users in the dataset regarding tags stored in $^{\mathrm{U}}\mathbf{A}$ derived according to Section 3.2, and determined the tags with highest influence. We assume that such characteristic tags have a value at least one standard deviation above the mean of $^{\mathrm{U}}\mathbf{A}$, leaving us with 3.46 tags per user.

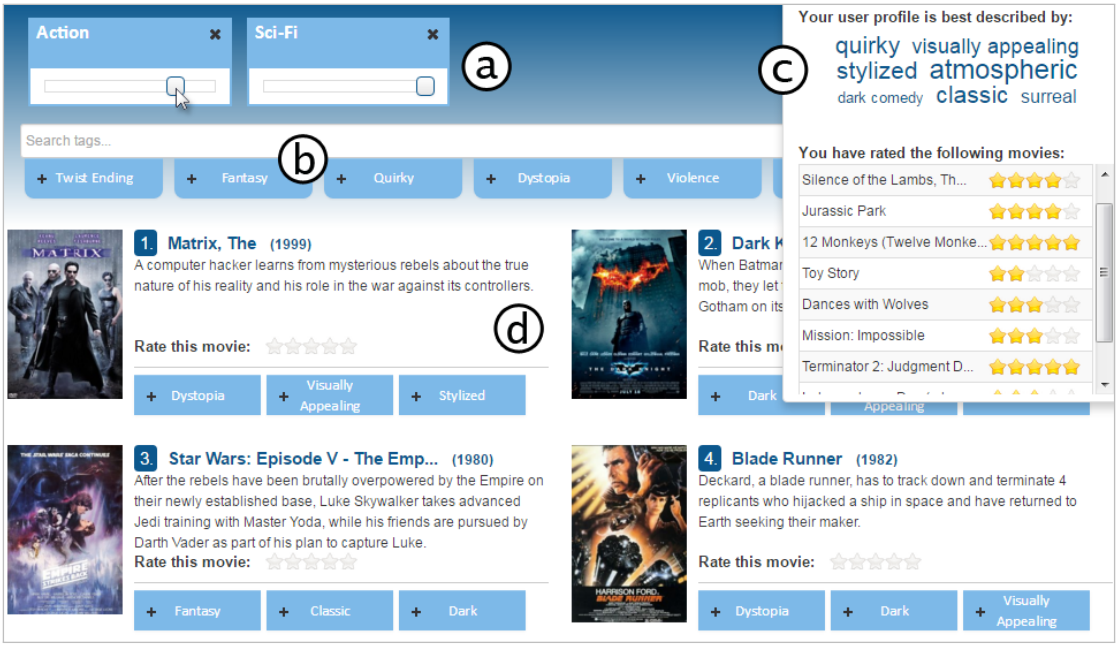

Figure 3: Screenshot of the prototype RS for the first user study: The current user has weighted the tags "action" and "sci-fi" (a), therefore receiving matching movie recommendations such as "Matrix" or "Star Wars" (d). The user can also search for other tags provided by the user community or get inspiration from the suggestions (b). Furthermore, the user's existing profile is explained by a tag cloud (c).

Afterwards, participants were presented with the now adjusted top-6 recommendations. Again, they had to settle on one movie, rate how satisfying each recommendation was, and fill in a questionnaire[14]. Note that this time, the questionnaire was complemented with items regarding the interaction process.

5.2.3. 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) and about newer movies (M=2.93, SD=0.98).

We conducted two-way repeated measures ANOVA to compare the effects of condition and point in time on specific dependent variables corresponding to our hypotheses. For the comparison between the three conditions, mean values and standard errors are reported in Table 2.

In the following, we elaborate on the statistical significance ($\alpha = .05$) of the differences found in these results. Moreover, we report differences with respect to point in time. Note that interaction terms between the two factors were never significant, so we omit presenting them. For post hoc comparisons, we used the Bonferroni test.

Perceived Recommendation Quality. Concerning the subjective assessment of recommendations, there was a statistically significant effect for condition, $F(2, 90) = 7.40$, $p < .001$, $\eta_p^2 = .14$, with large effect size. Post hoc tests indicated that the mean value for *TagMF*-Tags was significantly higher than for both, *TagMF*-Ratings, $p = .028$, and standard MF, $p < .001$. This confirms H1.

There was no significant difference regarding point in time, i.e. between before and after the respective interaction phase, $F(1, 45) = 0.02$, $p = .904$, $\eta_p^2 = .01$.

Mean Item Rating. With respect to ratings participants provided for each of the top-6 recommended items, we found differences to be similarly significant, $F(2, 88) = 11.19$, $p < .001$, $\eta_p^2 = .20$, again with large effect size. Movies in the *TagMF*-Tags condition received significantly higher ratings than in the two other conditions, *TagMF*-Ratings, $p = .025$, and standard MF, $p < .001$. As a consequence, we can eventually fully accept H1.

We found no significant effect with respect to point in time, $F(1, 44) = 0.02$, $p = .885$, $\eta_p^2 = .01$.

Table 2: Mean values and standard errors for the different conditions. Higher values indicate better results (*Choice Difficulty* and *Effort* are reversed accordingly), except for time values additionally reported.

	Standard MF		*TagMF*-Ratings		*TagMF*-Tags	
Construct	*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	3.19	0.15	3.03	0.15	**3.30**	0.15
	33.82 s	3.09	**28.41 s**	2.60	28.48 s	2.37
Effort	3.77	0.13	**3.84**	0.10	3.64	0.11
	165.54 s	16.9	224.96 s	20.05	194.41 s	19.21
Transparency	3.20	0.15	3.41	0.15	**3.73**	0.13

[14]For each condition, the dependent variables were thus assessed at two different points in time, i.e. before and after the respective interaction phase.

Choice Satisfaction. Regarding satisfaction with the movie participants finally selected from the set of recommendations, we found statistical evidence for differences between conditions, $F(2,90)=4.72$, $p=.011$, $\eta_p^2=.10$, with medium effect size. Post hoc tests indicated that the mean value for *TagMF*-Tags was significantly higher than for standard MF, $p = .009$, which confirms H2. No differences were found between *TagMF*-Ratings and other conditions.

Furthermore, we found a significant difference regarding point in time, $F(1,45)=5.07$, $p=.029$, $\eta_p^2=.10$, with medium effect size. Before interaction phases ($M=4.28$, $SE=0.10$), participants were more satisfied with their selected movie than afterwards ($M=4.02$, $SE=0.11$).

Choice Difficulty. We objectively operationalized the difficulty to decide as the total time participants spent for settling on a movie they would actually like to watch from the shown top-6 recommendations. The within-subjects main effect yielded significant differences with medium effect size for condition, $F(2,88)=5.34$, $p=.006$, $\eta_p^2=.11$. Participants took significantly more time with standard MF compared to *TagMF*-Ratings, $p=.015$, and *TagMF*-Tags, $p = .050$. The difference between the two *TagMF* conditions was not significant.

Participants decided more quickly after the interaction phases ($M=25.81$ sec, $SE=2.31$) than before ($M=34.66$ sec, $SE=2.88$), with significant difference and large effect size, $F(1,44)=28.03$, $p<.001$, $\eta_p^2=.39$.

In addition, we specifically asked participants how difficult it was to choose a movie[15]: With respect to their subjective perception, we neither found a significant effect for condition, $F(2,90)=1.20$, $p=.307$, $\eta_p^2=.03$, nor point in time, $F(1,45)=1.60$, $p=.212$, $\eta_p^2=.03$. Overall, we can thus only partly accept H3.

Effort. Concerning total time participants spent in the different conditions for the interaction phase, we found a significant effect using a one-way ANOVA, $F(2,90)=3.34$, $p=.040$, $\eta_p^2=.07$, with medium effect size. On average, participants needed significantly more time in the *TagMF*-Ratings condition compared to standard MF, $p=.040$. No differences were found in other pairwise comparisons.

However, although the interaction phase in both *TagMF* conditions was at least marginally longer, we found no significant differences with respect to perceived interaction effort[15], which we assessed after each interaction phase and analyzed using a one-way ANOVA, $F(2,90)=1.40$, $p=.253$, $\eta_p^2=.03$. Overall, this confirms H4.

Transparency. Once again using a two-way repeated measures ANOVA, we noted a significant effect of condition on transparency, $F(2,90)=6.22$, $p=.003$, $\eta_p^2=.12$, with medium to large effect size. Results from standard MF were perceived less transparent than from *TagMF*-Tags, $p=.003$, which confirms H5. No differences were found between *TagMF*-Ratings and other conditions.

Moreover, no significant effect was found for point in time, $F(1,45)=0.01$, $p=.948$, $\eta_p^2=.01$.

Usability. Regarding the two variants of our prototype RS, a paired t-test ($t(45)=4.15$, $p<.001$) indicated that participants generally preferred the variant that integrated interactive features based on *TagMF* ($M=3.76$, $SD=1.02$) over the one that used standard MF ($M=2.83$, $SD=1.00$). Some participants, for instance, explicitly stated that they "do not want to use only star ratings, but rate several aspects, so that the system can better recommend movies" and "really liked the tag selection with the sliders".

More specifically[16], usability of the prototype variant that supported interaction via tags was rated as "good" with a SUS score of 78. Values between 0.95 and 1.96 on the different subscales of the UEQ were equally promising. 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 positive assessment of interface adequacy ($M=4.13$, $SD=0.48$). Overall, the variant was rated to be particularly useful with no ($M=3.78$, $SD=0.99$) or only a vague ($M=3.89$, $SD=1.02$) search goal in mind. In contrast, but as expected, participants found it less suitable when they already knew their search direction ($M=2.52$, $SD=1.50$).

5.2.4. Structural Equation Modeling

Since we were particularly interested in differences between conditions in cold-start situations where the system must deal with high uncertainty, we used SEM (see Section 2.4) to further investigate the effects of using different recommender algorithms and methods for eliciting initial preferences on subjective system aspects and user experience.

Based on the framework for user-centric evaluation of RS proposed in [17] (see Section 2.4), we defined algorithm (*Standard MF vs. TagMF*) and initial preference elicitation method (*Ratings vs. Tags*) as *Objective System Aspects* (OSA) that cannot be influenced by the user. We considered *Perceived Recommendation Quality* and *Transparency* as *Subjective System Aspects* (SSA) representing user perception of OSA. SSA are seen as mediating variables between OSA and user experience [17, 16]. User experience is known to be substantially affected by underlying algorithms and preference elicitation methods [e.g. 76, 17, 21, 77, 63, 16, 78]. In light of this, we assumed user experience and interaction behavior to be influenced through changes regarding SSA, when using, for example, a novel means for eliciting initial preferences such as selecting tags according to Section 4.1. Consequently, we

[15] Note that higher values indicate better results.

[16] Note that we only asked specific questions regarding usability for the *TagMF* variant to reduce participants' workload in the within-subject design. Besides, interaction in the other variant was limited to rating items, minimizing the need for a separate evaluation.

included *Choice Satisfaction* as an indicator of *User Experience* (EXP), and complemented the more general perceived quality of the set of top-6 recommendations by capturing *Interaction Behavior* (INT) in form of the specific rating feedback for the individual movies, i.e. *Mean Item Rating*. In addition, we took personal characteristics into account to deduce assumptions about the influence of different dispositions. In line with the underlying framework, we assumed attitude and behavior concerning the varied system aspects to be affected by certain *Personal Characteristics* (PC) such as *Domain Knowledge* and *Trust in Technology*.

We set up a first theoretical model (Figure 4), yielding a good fit with the data ($\chi^2(7) = 8.246$, $p = .311$, $CFI = .995$, $TLI = .989$, $RMSEA = .032$). It explains a large amount of variance regarding our dependent variables *Choice Satisfaction* ($R^2 = .408$) and *Mean Item Rating* ($R^2 = .698$), as well as about 20 % of our considered mediator *Perceived Recommendation Quality* ($R^2 = .208$).

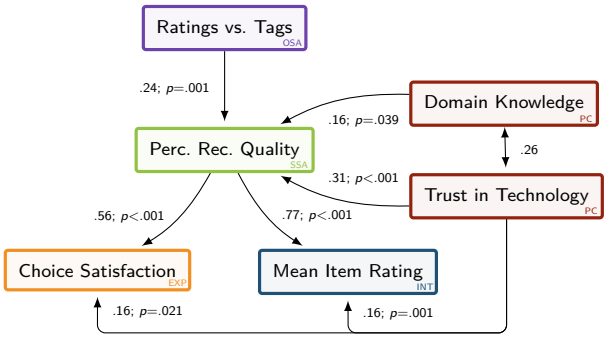

Figure 4: Path model for comparing the influence of initial preference elicitation via ratings or tags. On the edges, standardized regression weights as well as corresponding p-values are displayed.

Direct effects of varying the algorithm (*Standard MF vs. TagMF*) between conditions were not significant for any dependent variable or the mediator. Thus, this OSA was eventually not integrated in our model. The method for initial preference elicitation (*Ratings vs. Tags*) seems in contrast to account for a significant explanation of *Perceived Recommendation Quality*. Regarding personal characteristics, *Domain Knowledge* shows a meaningful influence only on *Perceived Recommendation Quality*, but *Trust in Technology* on all dependent variables.

The mediator *Perceived Recommendation Quality*, an overall subjective assessment, seems to be a strong predictor for both more specific variables, *Choice Satisfaction* and *Mean Item Rating*. Further analysis shows that *Perceived Recommendation Quality* appears to completely mediate the otherwise significant predictive power of varying the initial preference elicitation method (*Ratings vs. Tags*).

In view of our hypotheses, we aimed at further clarifying the role of participants' understanding of recommendations in cold-start situations (H5). Thus, we integrated *Transparency* as additional mediator in a second model (Figure 5). Overall, this model, which again fits the data well ($\chi^2(12) = 13.669$, $p = .322$, $CFI = .995$, $TLI = .989$, $RMSEA = .032$), explains a large proportion of variance regarding *Choice Satisfaction* ($R^2 = .401$), *Mean Item Rating* ($R^2 = .693$) and *Perceived Recommendation Quality* ($R^2 = .523$). Moreover, it achieves a reasonable amount of explained variance with regard to *Transparency* ($R^2 = .234$).

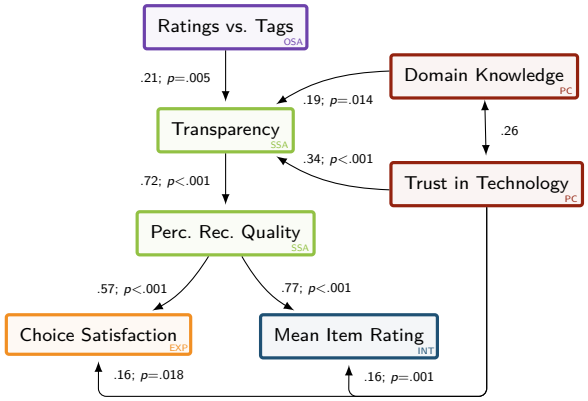

Figure 5: Path model for comparing the influence of initial preference elicitation via ratings or tags, mediated by transparency. On the edges, standardized regression weights and p-values are displayed.

The second proposed model shows that the predictive power of *Perceived Recommendation Quality* on the dependent variables observed in the first model obviously still holds. However, there are significant shifts of relations between the variables due to integrating *Transparency*: *Transparency* seems to be a substantial causal factor for *Perceived Recommendation Quality*, which in turn acts as a complete mediator for the effects on our more specific dependent variables. In fact, *Transparency* itself seems to be a regressor fully mediating the direct effect of varying the initial preference elicitation method (*Ratings vs. Tags*) on *Perceived Recommendation Quality* found in the first model. This confirms H5 even for the special case of cold-start. Besides, *Transparency* appears to be a partially mediating variable for the personal characteristics *Domain Knowledge* and *Trust in Technology*.

5.2.5. Discussion

In the first user study, the variant of our prototype system that relied on *TagMF* received significantly higher scores with respect to a number of variables related to subjective system aspects and user experience. Even in cases differences were found not significant between standard MF and *TagMF*-Ratings, results in the condition based on the extended variant tended to be better[17]. Regard-

[17] In all conditions, some scores were not as high as expected. We assume this to be due to the dataset (only movies released before 2008) and our particular sample of participants (more females, rather young, average domain knowledge). Qualitative answers to the open-ended question as well as better results in the second user study (newer dataset, more homogeneous sample) support this assumption.

ing assessment before and after interaction phases, perceived recommendation quality and transparency did not differ significantly. Since interaction terms of condition and point in time were never significant, we deduce that this applies to all conditions. Satisfaction with the chosen movie was even decreased after interaction phases[18], but this can be justified by examining typical user behavior: Participants rated movies they knew and liked already during these phases. Consequently, the result set changed a lot, eventually comprising items which might be not as easy to assess at first sight. One participant explicitly mentioned that the recommendation set "would have better fitted his or her taste when movies he or she rated highly had not been removed". Still, scores related to recommendation quality (H1), choice satisfaction (H2) and transparency (H5) seem very promising, and more importantly, higher in both *TagMF* conditions.

In real-world scenarios, initial preference elicitation—here performed as a preliminary task—would be part of actual system use. In this context, the significant differences between conditions before the interaction phase (with *TagMF* being superior) suggest that the few interaction steps initially taken, i.e. selecting a small number of tags up front, are already sufficient to improve user experience of typical RS. In particular, transparency in the *TagMF* conditions was rated better even at the beginning, although participants did not know that results were based exclusively on few tags. Thus, considering additional content information according to our method seems to help users implicitly when judging recommendations independent of later interaction (H5).

Because of these findings, we further examined the role of transparency at cold-start by using SEM. Our first proposed model indicated that selecting tags instead of rating items to elicit initial preferences significantly improves perceived recommendation quality (H1). Including transparency into the second model increased the amount of variance explained by the entire model concerning perceived recommendation quality from 21 % to 52 %. With a high standardized regression weight, transparency appears to be a substantial predictor for perceived recommendation quality. In turn, varying the preference elicitation method significantly contributes to explaining transparency (see Figure 5). The second model further shows that the effect on perceived recommendation quality found in the first model is actually fully mediated by transparency. Apparently, relying on *TagMF* leads to more comprehensible results (H5), which are consequently perceived to be of higher quality, ultimately also increasing participants' satisfaction with their chosen item (H1, H2). We deduce that user-generated tags import semantics into the result set which are more natural to understand than a meaning derived from recommendations calculated exclusively based on typical user-item interaction data. Our qualitative inspection of the factor space supports this (Section 5.1.3). In summary, the significant influence of initial preference elicitation method emphasizes that selecting tags according to the application possibility described in Section 4.1 instead of rating items up front is a promising means to alleviate the cold-start problem in model-based CF systems (RQ1a).

As a side note, while recommendation quality was indeed the main predictor for choice satisfaction and individual rating feedback, also personal characteristics had an impact. For instance, by increasing transparency, our method seems particularly useful for users with little domain knowledge, as it becomes easier to comprehend why certain items are recommended. The influence of trust in technology was however only partially mediated via transparency. Personal characteristics thus might alter the way perceived quality is translated into numerical ratings: Users whose trust in technology is low are likely to provide lower ratings in a more technically-oriented system. This poses another argument for using more natural ways to interact with CF systems than (re-)rating single items.

System usability was assessed overall very positively for the prototype variant based on *TagMF*[16]. Some participants had specific suggestions (e.g. "full text search should be integrated") or complaints (e.g. "movies cannot be excluded from the results without rating them") with regard to system functionalities. However, these qualitative comments addressed rather general usability issues beyond the scope of our research, and were, in particular, not exclusively related to the variant that supported interaction via tags. When asked specifically, participants in general preferred this variant. While this might be a reason why they spent more time in the two corresponding conditions (significantly or at least tendentially longer interaction phases), the richer interaction possibilities may account for this finding as well. Moreover, participants had to get used to the novel mechanisms introduced by *TagMF* while they were likely more familiar with conventional rating-based interfaces. Either way, perceived interaction effort did not differ significantly, so that we can accept H4.

In summary, our interactive approach realized by applying *TagMF* seems valuable for improving transparency of recommendations as well as providing users with extended possibilities to control the recommendation process and to adapt the results towards their current interests when relying on an existing long-term profile, thus validating the application possibility described in Section 4.2 (RQ1b).

Lastly, with our study, we for the first time confirmed that enhancing model-based CF with additional information is beneficial with respect to subjective perception of recommendation quality, which previously has only been observed in terms of offline performance (see Section 2.2).

[18] In terms of choice difficulty operationalized as the time spent for settling on a movie, the point in time also had a significant effect. However, this was expected as it is likely that participants already decided for an item during interaction phases, and were therefore able to choose faster when asked to choose a movie afterwards. Note that the subjective perception did not differ significantly.

For cold-start situations, SEM however showed no significant difference in this regard when varying the algorithm. This is generally in line with recent research stating that different or objectively more accurate recommenders do not necessarily produce better results from a user perspective [5, 6, 7, 63]. Although it may achieve high accuracy scores, a list of items detached from a superordinate context might not be satisfactory for users. Instead, the recommendation set should exhibit some kind of inner consistency, which in our case is reached through establishing a relationship between latent factors as derived by MF and user-generated tags. Recommendations thus seem to refer to each other implied by the easy-to-understand semantics of tags. Consequently, using *TagMF* positively affects transparency by building a meaningful context, thereby in turn improving perceived recommendation quality. In accordance with this, participants needed significantly less time to choose a movie from the recommendations in the respective conditions (H3). Beyond that, our study showed that compared to a typical automated RS based on ratings, also other subjective aspects related to user experience benefit equally from considering additional information according to our method (RQ2a).

5.3. Empirical User Study II

We performed the second user study with the goal of investigating the influence latent knowledge has on the recommendation process from a user perspective. In this regard, we wanted to focus on the comparison against an interactive RS that relies on user-generated content alone, and to examine the value of *TagMF* for implementing critiquing.

5.3.1. Goals

First, we aimed at validating another application possibility of *TagMF*: For evaluating the option to interactively critique a recommended item by means of user-generated tags in a model-based CF scenario (RQ1c), we implemented it according to Section 4.3 in our web-based prototype movie RS. Next, since we were interested in examining the impact using a latent factor model that integrates additional information has on the subjective assessment of system aspects, and thus on user experience, when compared to a purely tag-based interactive approach (RQ2b), we formulated the following hypotheses contrasting this baseline and *TagMF*:

- **H1:** *TagMF* improves perceived quality of recommendations.
- **H2:** *TagMF* improves satisfaction with the item chosen from the recommendations.
- **H3:** *TagMF* decreases difficulty to choose an item.
- **H4:** *TagMF* decreases perceived interaction effort.
- **H5:** *TagMF* leads to more diverse recommendations.
- **H6:** *TagMF* has no negative impact on transparency.
- **H7:** *TagMF* improves perceived quality of critiquing.

5.3.2. Method

The study was designed as an experiment under controlled conditions. We had 54 participants (37 female) with an average age of 27.89 ($SD = 10.30$), a small majority of them students (57 %). To interact with the prototype RS and to answer questionnaire items, they used a common web browser running on a desktop PC with a 24" LCD (1920×1200 px resolution). Next, we describe the prototype system, the procedure, and the used questionnaire, in more detail.

Prototype. Figure 6 shows the web-based prototype movie RS we developed for the second user study. We again set up two variants: One reimplementing the method behind *MovieTuner* [18], with an interface resembling a typical critique-based RS (see Section 2.1), in particular, the integration of *MovieTuner* in the MovieLens platform [18]. This purely tag-based variant served as a baseline to test our hypotheses. The other variant with nearly identical interface was implemented using *TagMF*. Here, to validate this application possibility, we integrated the interactive critiquing process as described in Section 4.3.

Concretely, the interface is structured as follows: At the top (a), the critiquing area comprising tags used as dimensions to critique the currently recommended item is displayed. As in the *MovieTuner*, these tags generated by the user community are automatically shown by the system based on the method described in [18]. This method considers tag utility, popularity and diversity to determine a set of tags that is particularly meaningful for critiquing the current item. The only requirement is availability of item-related tag relevance scores, i.e. our given matrix ${}^{\mathbf{i}}\mathbf{A}$ (see Section 3.1). However, in the variant relying on *TagMF*, we additionally exploit that user-item interaction data is available as usual in CF systems, and blend this set together with a set of tags reflecting the current user's specific interests. Concretely, we replace half of the presented critique dimensions with tags scoring highest for this individual user, thereby personalizing the critiquing area. This is possible with *TagMF* as we also know user-related tag relevance scores for all available tags provided by the user community, i.e. our derived matrix ${}^{\mathbf{u}}\mathbf{A}$ (see Section 3.2). Either way, each tag is accompanied with radio buttons allowing users to critique the currently recommended item (details for this movie are shown on demand when hovering its title), i.e. requesting new suggestions with less, equal, or more relevance with respect to the corresponding tag (we implemented critiquing for the two system variants according to the respective method, as described below).

Moreover, users can search and manually choose tags as additional critique dimensions using the input field underneath (b), supported by autocompletion. As in the first user study (see Section 5.2.2), all available tags come from the underlying dataset (see below) and are generated by other users. Yet, since *TagMF* can be used with any set of tags, it would indeed be possible to also create new tags

in a real-world system.

In the *TagMF* variant of our prototype, the user profile is additionally presented in a dialog (c) similar as in the prototype for the first user study (see Section 5.2.2). The rest of the screen shows the top-9 recommendations. Note that in contrast to standard critique-based RS, and thus to the tag-based prototype variant, these recommendations in the *TagMF* variant rely on both the user's situational needs, i.e. critiques applied to the currently recommended item, as well as his or her long-term profile based on historical preference data as it is customary in CF systems. Each recommendation (d) is displayed together with the 3 tags most relevant to the respective movie (these tags may be selected as critique dimensions as well), and a button that may be used to choose this movie as a new item to critique, i.e. to start a new cycle in the critiquing process.

For calculating recommendations in the *TagMF* variant, we again used a Stochastic Gradient Descent MF algorithm as point of departure. As a result of the offline experiments reported in Section 5.1, we used 20 factors, 30 iterations, and set $\lambda = .001$. Moreover, we used the 50 most popular user-generated tags from the underlying dataset as additional training data, and integrated the resulting model-based CF recommender with critiquing as described in Section 4.3. For generating recommendations and integrating the critiquing process in the other variant of our prototype system, i.e. the one exclusively based on tags, we reimplemented the method behind *MovieTuner* as proposed in [71, 18, 79]. For this, we again relied on the 50 most popular tags. According to prior testing, we chose the *linear-sat* metric for computing critique satisfaction. Further parameters are set as suggested in the literature. For item data, associated ratings and tag relevance scores, we used the same intersected dataset based on MovieLens 20M and MovieLens Tag Genome dataset as in the offline experiments (see Section 5.1.1). While we thus relied on scores precomputed as described in [71] based on user-generated tags from the underlying dataset, one would in a real-world scenario indeed use tags provided by users of the system at hand to calculate these scores.

Questionnaire and Log Data. As in the first user study, the questionnaire participants had to fill in was primarily based on the pragmatic evaluation procedure for RS proposed in [72], containing items related to subjective system aspects and user experience (see Section 5.2.2). Concretely, we assessed *Perceived Recommendation Quality*, *Choice Satisfaction*, *Choice Difficulty*, *Effort* and *Diversity* by means of items from this framework. We used an item from [73] to assess *Transparency* of recommendations. Regarding the *Critiquing*, we selected items from [18].

Again using items from [73], we assessed the overall satisfaction of participants with the respective prototype variant as well as the interface adequacy. In addition, we applied *System Usability Scale* (SUS [74]) and *User Experience Questionnaire* (UEQ [75]). We gathered data about demographics and familiarity of participants with the movie domain. Apart from UEQ (7-point bipolar scale), all items were assessed on a positive 5-point Likert-scale (1–5). We also collected qualitative feedback: An open-ended question asked participants to report suggestions and complaints. Finally, we logged user interaction behavior and measured task times.

Procedure. First, participants were asked to complete a preliminary task that served to elicit an initial set of preferences. For this, movies were presented one after the other based on popularity and entropy as proposed in [80]. Items were separated into blocks of 25, and then shuffled to eliminate sequence effects. Unknown movies could be skipped. After participants rated 10 movies, this feedback was used to initialize a standard factor vector using online-updating (again implemented according to [49]) and to subsequently generate recommendations using *TagMF*: The top-15 results were presented in form of a list that could be expanded up to a maximum of 30 movies. Participants had to choose one movie out of these recommendations which they should know, and would find interesting as a starting point for a succeeding critiquing process.

Next, participants were assigned in counter-balanced order in a between-subject design to one of the two following conditions that correspond to the two system variants (yielding 27 participants per condition):

Tag-based: Tag-based method with recommendations based on similarity and critique distance to the currently recommended item in terms of tag relevance, implemented according to the method behind *MovieTuner* [71, 18, 79]. Critique dimensions were shown based on item-related tag relevance scores. Participants could interactively select tags, apply critiques, and switch the critiqued item.

***TagMF*:** Tag-enhanced MF with recommendations based on user profile (i.e. derived factor vector) and currently recommended item as described in Section 4.3. Critique dimensions were suggested based on item-related as well as user-related tag relevance scores. Interaction was equivalent to the other condition.

In both conditions, participants were initially presented with an item representing the starting point for the critiquing process (see task descriptions below) as well as the top-9 recommendations generated according to the underlying method. Note that the only visible difference in the interface of the two prototype variants was availability of the dialog showing the user profile (see Figure 6). In the background, however, the way critique dimensions were selected and recommendations were generated differed as described above. Either way, participants had to interact with the respective system variant, i.e. apply critiques and switch the currently critiqued item, in order to refine the set of 9 recommendations and to fulfill the following tasks:

1) Participants were asked to find a movie that fits their personal preferences and they would actually like to watch. As a starting point, the movie chosen after the preliminary task was shown. Recommendations

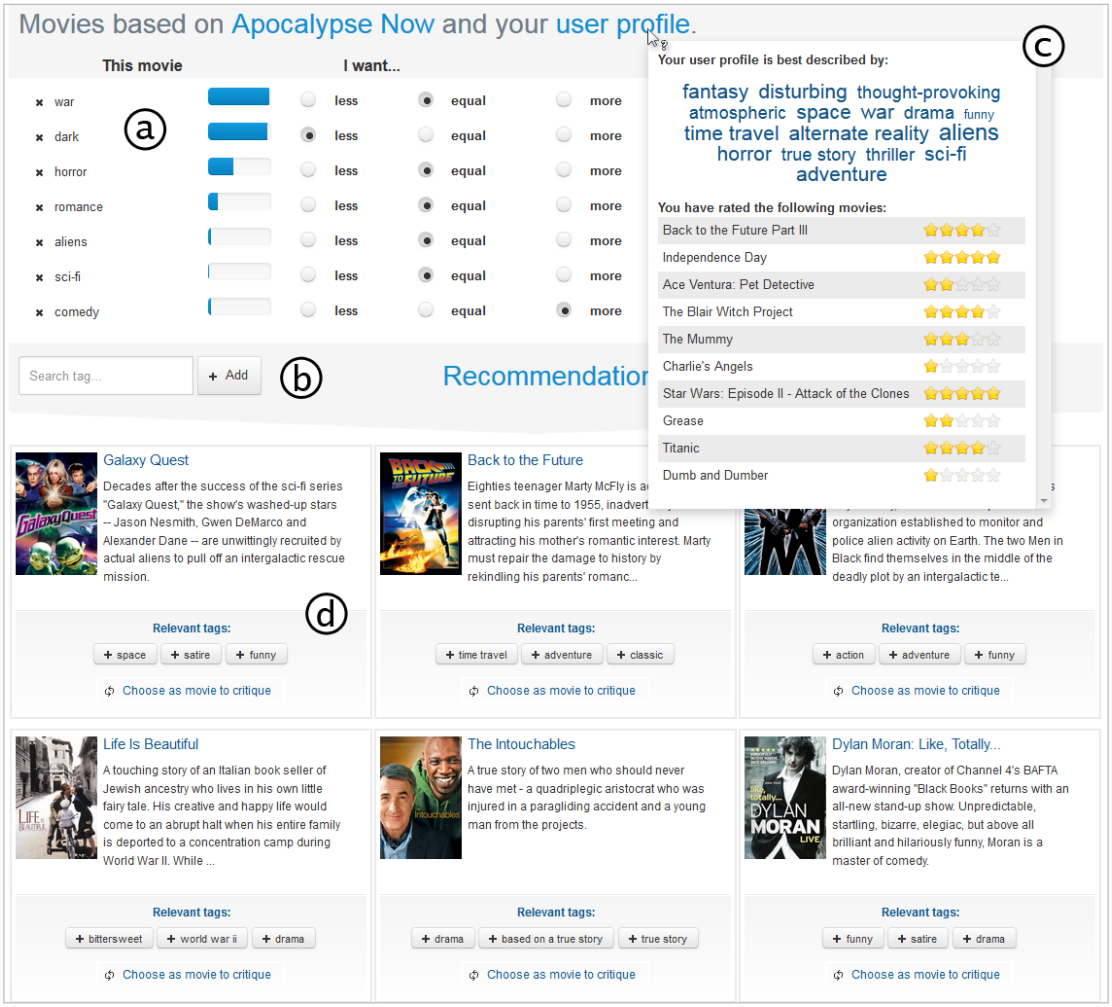

Figure 6: Screenshot of the prototype RS for the second user study: A user whose profile is shown in the dialog (c) has applied a critique (a) to the currently recommended movie "Apocalypse Now" using the tags "dark" and "comedy". As a consequence, recommendations that fit to the critique and to his or her long-term interests are shown (d). To add further critique dimensions, the user can also search for tags provided by other users (b).

based on the currently critiqued item, and in the *TagMF* condition additionally on the preferences of the current participant, i.e. the factor vector learned by means of the 10 ratings elicited up front.

2a) Participants were asked to find a movie that they would like to watch when going out on a date with someone. Thus, they were required to not only take their personal preferences into account, but in addition interests of the fictitious date (which were not explicitly given). As a starting point, the movie chosen after the preliminary task was shown. Recommendations were generated as in the previous task.

2b) Participants were asked to find an adequate movie for the given situation that an adult horror movie fan wants to watch a movie together with a 9-year-old child. Thus, they were required to assume a high interest in horror movies while taking the interests of the child into account (which were not explicitly given). As a starting point, we selected a representative horror movie. Recommendations based on the currently critiqued item, and in the *TagMF* condition additionally on an artificial profile we created by training a factor vector with ratings typical for a horror movie enthusiast.

Task 2a and 2b were presented in random order. All tasks were finished by participants at their own discretion.

After each task, participants were first asked to choose the movie they found most suitable for the given task from the final set of top-9 recommendations. Second, they had to rate their satisfaction with each of the recommended items on a 5-point Likert-scale (1–5). Finally, participants were asked to fill in the questionnaire as described above.

5.3.3. Results

Participants of the second user study reported average knowledge about movies (M=3.02, SD=0.87). The movie chosen initially as a starting point after completing the preliminary task was rated very positively (M=4.65, SD=0.68), while most participants had seen it (94 %).

For the directed hypotheses, we conducted one-tailed t-tests (using a significance level of $\alpha = .05$) to compare the two conditions in terms of corresponding dependent variables. In contrast to the first user study with a repeating task in a within-subject design, the different nature of the tasks in the second user study made a comparison between tasks rather meaningless. Instead, we were specifically interested in individual results per task. Thus, we omit reporting repeated-measures variance analyses, but present the results with respect to subjective system aspects and user experience separately for each task in Table 3.

Table 3: t-test results with mean values and standard deviations ($df = 52$, except for † (48.36) and ‡ (45.51) adjusted due to unequal variances) comparing the conditions with respect to subjective system aspects and user experience (* indicates significance at 5 % level; d represents Cohen's effect size value). Higher values indicate better results (*Choice Difficulty* and *Effort* are reversed accordingly).

	Tag-based		*TagMF*				
Construct & Task	*M*	*SD*	*M*	*SD*	*T*	*p*	*d*
Perc. Rec. Quality							
Task 1	3.67	0.84	**4.20**	0.67	2.59	.006*	0.71
Task 2a	3.87	0.93	**4.19**	0.86	1.30	.100	0.35
Task 2b	3.26	0.81	**4.02**	0.88	3.29	.001*	0.90
Mean Rating							
Task 1	3.61	0.55	**3.83**	0.66	1.32	.097	0.36
Task 2a	3.45	0.49	**3.86**	0.57	2.75	.004*	0.76
Task 2b	3.27	0.55	**3.65**	0.64	2.30	.013*	0.63
Choice Satisfaction							
Task 1	4.59	0.50	**4.78**	0.64	1.18	.121	0.32
Task 2a	4.56	0.64	**4.81**	0.48	1.68†	.050*	0.46
Task 2b	4.00	0.83	**4.52**	0.64	2.56	.013*	0.70
Choice Difficulty							
Task 1	**3.59**	1.01	3.22	1.28	-1.18	.122	-0.32
Task 2a	**3.37**	1.15	3.33	1.33	-0.11	.457	-0.03
Task 2b	2.89	1.09	**3.19**	1.30	0.91	.184	0.25
Effort							
Task 1	3.98	0.60	**4.06**	0.80	0.39	.351	0.11
Task 2a	3.89	0.87	**4.09**	0.75	0.92	.180	0.25
Task 2b	3.46	0.63	**3.72**	0.94	1.19‡	.121	0.32
Diversity							
Task 1	3.67	0.92	**4.07**	0.83	1.71	.047*	0.47
Task 2a	3.89	0.89	**4.19**	0.62	1.42	.082	0.39
Task 2b	3.81	0.79	**4.11**	0.75	1.42	.082	0.39

Perceived Recommendation Quality. Concerning perceived quality of recommendations, we found a statistically significant effect for condition in Task 1 and Task 2b. Mean values for *TagMF* were significantly higher than in the tag-based condition. Note that effect sizes were medium to large, or at least small to medium for Task 2a. Overall, this confirms H1.

Mean Item Rating. Individual ratings participants provided for the top-9 recommended items in the *TagMF* condition were found significantly higher in Task 2a and 2b, with medium to large effect sizes. Although there was no significant difference in Task 1, the ratings given in the *TagMF* condition were on average higher than in the tag-based condition, with small to medium effect size. Thus, we can eventually fully accept H1.

Choice Satisfaction. Participants in the *TagMF* condition were more satisfied with the movie chosen from the set of top-9 recommendations in all tasks. For Task 2a and 2b, we even found statistical evidence for differences between the tested conditions, with medium to large effect sizes. Overall, this confirms H2.

Choice Difficulty. Regarding the subjective assessment of the difficulty to choose one item from the set of movies eventually recommended[15], we found no significant differences between conditions. In two comparisons, the tag-based variant received marginally better results, but with rather small effect sizes. Nevertheless, we have to reject H3.

Effort. Interaction effort was perceived slightly better in the *TagMF* condition[15]. However, we could not observe significant differences. This is reflected in task times, which likewise did not differ between conditions: Task 1: $t(52) = 1.24$, $p = .111$, $d = 0.34$; Task 2a: $t(52) = -0.25$, $p = .401$, $d = -0.07$; Task 2b: $t(52) = 1.06$, $p = .147$, $d = 0.29$. Overall, while *TagMF* at least tended to get better subjective results in all tasks, we have to reject H4.

Diversity. Participants rated the diversity of the set of recommendations generated by *TagMF* higher than participants in the tag-based condition. With medium effect sizes for all tasks, we even found a significant difference between the two conditions in Task 1. Overall, this confirms H5.

Transparency. Once after completing all tasks, we asked participants how they perceived the transparency of recommendations. They provided better scores in the *TagMF* condition ($M = 4.22$, $SD = 0.89$) than in the tag-based one ($M = 4.15$, $SD = 0.82$). Admittedly, the effect size was small ($d = 0.09$), and with a two-tailed t-test we found no evidence for a significant difference ($t(52) = 0.32$, $p = .752$). This, however, confirms H6.

Critiquing. Regarding the critiquing process, and in particular, the tags we used as critique dimensions in our prototype, a MANOVA aggregating several questionnaire items taken from [18] indicated no significant difference between conditions, $F(12, 41) = 0.68$, $p = .761$, $\eta_p^2 = .17$. Table 4 shows the individual results for these items, which were assessed once, after participants completed all tasks.

Overall, we found that participants understood the critique dimensions and their effect on the results. Moreover, they liked to apply critiques in form of user-generated tags to influence the recommendation process. Considering qualitative feedback, one participant, for instance, answered to the open-ended question that it was "clear and straight-forward to point the system in the direction of movies he or she would like to watch". However, others commented that it "would have been helpful to see a list of all tags as it was difficult to come up with suitable ones" (note that autocompletion was provided) and that

Table 4: t-test results with mean values and standard deviations ($df = 52$) comparing the conditions with respect to the tags used as critique dimensions (d represents Cohen's effect size value). Higher values indicate better results.

	Tag-based		*TagMF*				
Questionnaire Item	*M*	*SD*	*M*	*SD*	*T*	*p*	*d*
The tags made sense to me	4.22	0.75	**4.48**	0.75	1.27	.106	0.35
The tags helped me learn about the movie	4.00	0.73	**4.26**	0.76	1.27	.105	0.35
I like having the ability to specify critiques	4.52	0.64	**4.67**	0.68	0.82	.207	0.23
Movies displayed in response to my critique made sense	3.67	1.04	**3.89**	1.12	0.76	.227	0.21

Table 5: t-test results with mean values and standard deviations ($df = 52$) comparing the conditions with respect to the UEQ subscales (d represents Cohen's effect size value). Higher values indicate better results on the 7-point bipolar scale.

	Tag-based		*TagMF*				
Subscale	*M*	*SD*	*M*	*SD*	*T*	*p*	*d*
Attractiveness	1.73	0.65	**1.99**	0.79	1.28	.206	0.35
Transparency	2.20	0.63	**2.43**	0.61	1.32	.194	0.36
Efficiency	1.70	0.62	**1.95**	0.63	1.47	.148	0.40
Control	1.22	0.58	**1.61**	0.75	2.13	.038*	0.58
Stimulation	1.36	0.66	**1.78**	0.80	2.09	.042*	0.57
Novelty	1.18	0.94	**1.34**	1.09	0.60	.550	0.16

they "missed a broader range of tags to select from". Still, the questionnaire results were very positive in both conditions, with mean values even being slightly higher for *TagMF*. In summary, in spite of the lack of significances (according to the one-tailed t-tests we conducted, see Table 4) and of high effect sizes, we can thus at least partly accept H7, especially considering the minor (and only algorithmic) differences between conditions with respect to the critique dimensions.

Usability. In line with our more specific hypotheses, we used a one-tailed t-test to analyze whether participants were in general more satisfied in the *TagMF* condition: Results indicated a higher satisfaction ($M = 4.48$, $SD = 0.75$) with the corresponding system variant than in the control group ($M = 4.11$, $SD = 0.80$), with significant difference and medium effect size ($t(52) = 1.75$, $p = .043$, $d = 0.48$). One participant in the *TagMF* condition, for example, explicitly stated the he or she "enjoyed using the system".

More concretely, usability of the two variants of our prototype system was rated equally "good", with a SUS score of 87 in the *TagMF* condition, and 84 in the other. A two-tailed t-test showed no significant difference ($t(52) = 1.12$, $p = .269$) and only a rather small effect ($d = .30$). This corresponds to the very positive assessment of interface adequacy in both the *TagMF* condition ($M = 4.44$, $SD = 0.57$) and the tag-based one ($M = 4.20$, $SD = 0.53$), without significant difference ($t(52) = 1.55$, $p = .128$) and medium effect size ($d = .44$). Regarding the UEQ, values between 1.34 and 2.43 on the different subscales were very promising for *TagMF*, as shown in Table 5. In particular, subscales for transparency and efficiency yielded "excellent" scores, and control was rated as "good". Overall, scores were inferior in the tag-based condition, with values in a range from 1.18 to 2.20. Efficiency was only rated as "good" and control as "above average". In terms of control and stimulation, two-tailed t-tests even indicated significant differences with medium effect size.

5.3.4. Discussion

In the second user study, we examined how *TagMF* can be exploited for integrating model-based CF with interactive critiquing, taking the critiques applied to the currently recommended item and, in contrast to typical critique-based RS, the user's existing long-term preference profile into account. First, we would like to draw attention on the very positive assessment of the movie chosen as a starting point after completing the preliminary task. Participants were asked to select this movie from the initial set of recommendations notably generated by means of *TagMF* in both conditions. The results corroborate findings from the first user study showing that our method indeed leads to very adequate suggestions (see Section 5.2.3).

After each of the main tasks, we obtained very promising results regarding perception of recommendation quality (H1). With exception of Task 2a, differences were significant[19]. When participants had to find movies fitting their personal interests, i.e. especially in Task 1, the value of *TagMF* for the critiquing process became even more apparent: The underlying CF model allows to consider preference profiles (i.e. user-factor vectors) learned over a potentially longer period of time via conventional preference elicitation. Thus, suggestions are more likely to correspond not only to critiques applied due to situational aspects of the search process, but to the user's general interests as it is typical for CF recommenders. This positive assessment of the recommendations is reflected in the scores for the more specific constructs, mean item rating and choice satisfaction, which are higher for *TagMF* in all cases, most often significantly (H1, H2).

The positive impact latent knowledge has on the critiquing process and resulting recommendations compared to when only (user-generated) content information serves as input data (as it is customary in critique-based RS), is also supported by other relevant variables related to subjective system aspects and user experience. For instance, while purely content-based approaches are known to tend to over-specialization [81], i.e. recommending similar items, we found significant or at least marginal improvements regarding diversity of recommended item sets due to using our method (H5). This is well in line with other works that propose to exploit latent factors to di-

[19] Potentially because it was harder for participants to determine whether recommended items fitted the goal of the task than in the two other tasks (as interests of the fictitious date had to be taken into account), the difference here was only marginal.

versify RS results or to address the "filter bubble" problem [e.g. 58, 60]. Concerning choice difficulty, we in contrast did not find a positive effect: The prototype variant that reimplemented the method behind *MovieTuner* even tended to make it easier to choose a movie from the set of recommendations[20]. As a consequence, while we assumed that taking long-term interests into account by means of *TagMF* would make it more easy to decide, we have to reject H3. However, usability assessment of the different system variants indicated no negative impact on user experience in general. As in the first user study, most usability-related comments were independent of the respective condition: In their qualitative feedback, participants wanted, for instance, to "directly search for movies" or to "exclude bad movies and keep good movies over several critiquing cycles". Consequently, we will address these more general aspects in future iterations of our prototype system, although they are actually more related to system use in real-world scenarios.

With respect to transparency, we found only marginal improvements due to using *TagMF*. Bearing in mind that in this case latent information comes into play, the results however even shed a positive light (H6): It would not have been surprising if the variant that exclusively relied on well understandable tag-based information had facilitated the comprehension of recommendations. In principle, the same applies to perceived effort and the more objective measurement, time spent for tasks. Yet, we assumed that considering the user's preference profile would have a positive effect on his or her efficiency when navigating through the information space. As the results were however only slightly better with our method, we have to reject H4.

Besides aspects related to recommendations, we investigated the effect of *TagMF* on the selection of the critique dimensions that served as a means to take participants' interests as well as task-related goals into account. Overall, participants expressed more positive feedback. Differences between conditions were not significant, but we blended together tags chosen according to our method with the ones selected based on item-tag relevance data to equal proportions, i.e. only 3 of the user-generated tags shown as critique dimensions were actually determined differently. This minor difference in the user interface, in combination with the between-subject design, might have diminished the effect of taking the CF profile into account. Effect sizes were still small to nearly medium, but it has to be noted that explanatory power was limited due to the number of participants. However, participants were confronted with the related questionnaire items only once, after completing all tasks. Thus, tasks where participants in addition to their personal preferences had to consider interests of others might have distorted the results: In these cases, personalized critique dimensions specifically tailored towards individual long-term interests by means of *TagMF* might indeed been less useful. The answers to the open-ended question support this assumption. For instance, one participant mentioned that it was "difficult to quickly change the direction of recommendations (from horror to comedy) in order to obtain movies for a 9-year-old". Yet, he or she explicitly added that "adapting to the user profile is, on the other hand, purpose of the system". Unfortunately, in contrast to the first user study (see Section 5.2.4), SEM did not lead to meaningful insights because of sample size and study design. As a consequence, we plan to further investigate how employing our method may affect the subjective assessment of critiquing, and thus user experience, with a larger number of participants. Overall, we can still at least partly accept H7.

In general, participants in the *TagMF* condition stated to be more satisfied than in the tag-based condition, with significant difference. Taken all together, enhancing model-based CF according to our method can thus be seen as a promising means to add interaction possibilities and to improve user experience. This validates that *TagMF* can be successfully applied as an extension as described in Section 4.3 to allow users critiquing a recommended item in this typically very restricted type of RS (RQ1c).

For specifically examining the value of learning a latent factor model that additionally integrates user-generated tags, the second user study was designed as a comparison with an interactive RS that similar to the well-known *MovieTuner* relied on an entirely tag-based model. As already outlined above, results were overall very positive: We observed that using *TagMF* led to significantly better recommendations in terms of subjective aspects such as perceived quality and diversity. The positive results are reflected in user experience, e.g. choice satisfaction, and in the higher average ratings provided for the items eventually recommended after finishing the critiquing process. On the other hand, particularly our usability evaluation yielded only slightly better results. Given sample size, the small visible differences between the prototype variants, and the potential confusion that might be induced by the latent factors in the *TagMF* condition, the absence of differences, however, already appears promising. Nonetheless, further investigation and larger user studies are required in this area. In summary, from a user perspective, considering additional content data according to our method yet seems beneficial in comparison to the wide range of interactive recommending approaches that solely rely on (user-generated) content information (see Section 2.1) and, as a consequence, cannot consider user profiles based on past user-item interaction data (RQ2b).

Finally, observing participants' behavior indicated that they valued that in the prototype variant based on *TagMF*, a tag cloud allowed to inspect their formerly opaque representation within the underlying model. The successful implementation of such a tag-based explanation in our pro-

[20] Note that in the first user study, we additionally assessed this variable objectively by measuring how long it took participants to settle on a recommended movie. Due to differences in study setup and task descriptions, it was only possible to assess this construct in a subjective manner for the second user study.

totype system according to Section 4.4 shows how additional information may be used in typical model-based CF systems for explaining existing long-term preference profiles (RQ1d). However, both user studies have not been focused on this aspect, making it subject of future work to more completely validate this application possibility. Concretely, we plan to conduct another empirical study to investigate actual comprehensibility of the tag cloud and to compare ours with other approaches that explain recommendations, in particular ones that use tags, e.g. [82].

6. Conclusions and Outlook

In this paper, we have introduced *TagMF*, a method that combines the benefits of latent factors derived by standard MF with the ones of user-generated tags. As discussed in Section 2, MF is an efficient means for generating precise recommendations and has been improved by several algorithmic advances in the last years, for instance, by enhancing the factor models with additional information. However, accuracy improvements as measured in retrospective offline experiments have not always contributed to user satisfaction to the same extent. Interactive recommending approaches, which have been shown to increase the level of user control and system transparency, in contrast, often use entirely different algorithms, thus being independent of the advantages provided e.g. by state-of-the-art model-based CF techniques.

Following our research questions posed at the beginning of this paper in Section 1, we have shown the value of additional content information such as tags when used in CF: Integrating item-related tag relevance data by means of *TagMF* allows to derive corresponding user-related tag relevance data (i.e. we do not require up front availability of tags describing the current user's interest, but instead infer this information) as well as tag-factor relations. This contributes to increased recommendation quality and simultaneously opens novel ways to extend typical automated RS with interactive techniques, thereby overcoming several of their widely discussed drawbacks. Users can be offered more control over the recommendation process, which is in contemporary real-world systems usually limited to (re-)rating single items. Concretely, the application possibilities of *TagMF* allow users to interactively adapt the set of items suggested as known from standard MF towards their current needs and goals through easily comprehensible tags—in cold-start situations (RQ1a), with an existing profile (RQ1b), and by critiquing a recommended item (RQ1c)—and to inspect their long-term preference profile with the aid of tag-based explanations (RQ1d).

The offline experiments we conducted corroborate that *TagMF* increases objective recommendation quality (see Section 5.1.2). Yet, it has to be noted that additional parameter tuning is necessary, i.e. the number of tags to be taken into account must be determined, and that model learning becomes more complex. On the other hand, a qualitative inspection performed in this context underlines that the method is able to reveal inherent meanings of the resulting, usually abstract latent factor models by incorporating the easy-to-understand semantics of user-generated tags (see Section 5.1.3). Still, one must also note that this data needs to be collected—or other datasources must be available—before being able to apply our method. Two quantitative user studies with an interactive web-based prototype movie RS served to validate the application possibilities proposed in Section 4, which are directly related to our research questions. While participants were not allowed to create tags themselves, we believe this involves no loss of generality as the well-known dataset we used consists of a very large set of tags generated by the user community of a similar system. Besides, in a real-world scenario, it would be possible to easily apply *TagMF* with any kind of additional data. Consequently, these studies together can be considered to constitute the first extensive empirical evaluation in RS research with respect to the use of additional information in CF.

The first user study presented in Section 5.2 confirmed for the first time a positive influence on the subjective assessment of system aspects, as well as on user experience in general. In particular, perceived recommendation quality and transparency benefited from the integrated *TagMF* model. As a consequence, participants were able to decide faster and were more satisfied with their chosen item. Interestingly, besides the fact that they liked the interaction via tags generally more, results were especially promising with respect to the elicitation of initial preferences. Apparently, integrating our method seems to be quite useful in cold-start situations, as selecting a small number of tags led to recommendations at least as good as rating a larger number of items ex ante. Using SEM, we further analyzed these findings, focusing on the role of transparency and the impact of different preference elicitation methods. In this way, overall, the first study allowed us to validate the application possibilities described in Section 4.1 and 4.2, referring to RQ1a and RQ1b, and being focused on a comparison with an automated rating-based MF system, to answer RQ2a.

The second user study presented in Section 5.3 shows the value of considering user-generated content in addition to latent knowledge in interactive recommending scenarios: The results emphasize that using *TagMF*, a personalized critique-based recommendation process can successfully be integrated as an extension to standard model-based CF systems. We again obtained positive feedback with respect to subjective system aspects, e.g. perceived recommendation quality and diversity, and regarding constructs related to user experience, e.g. choice satisfaction. Note that while the number of participants was limited in consideration of the study's between-subject design, the effect sizes in addition to statistical significances generally confirm the benefits of our method. Nevertheless, experiments with more participants would be required to reach significance more often and to make even stronger claims. Yet, especially given the minor differences between the condi-

tions, we believe that the current results already provide sufficient evidence in favor of our method. Overall, the second study therefore allowed us to validate the application possibility described in Section 4.3, referring to RQ1c, and to complement the user-centric evaluation against state-of-the-art recommending approaches by a comparison with a purely tag-based interactive recommender, thus answering RQ2b.

In summary, enhancing a recommender according to our method appears to be a promising means to provide additional interactive features in today's automated systems and to increase their transparency. We successfully addressed RQ1 by describing several application possibilities of *TagMF*, which we validated in our user studies. Qualitative comments of participants (see Section 5.2 and 5.3) suggest that there are indeed usability-related aspects of our prototype system that could be improved. Consequently, although more related to real-world use, these issues are of interest for future work and will then be further investigated by means of, among others, qualitative methods. In general, we still received very positive results with respect to usability, and in particular the novel interactive features that can be integrated in CF systems by using *TagMF*. In this context, it has to be noted that although it could have been expected, the extended interaction mechanisms had no negative influence on e.g. perceived effort. With regard to RQ2, the user studies allowed us to investigate the effect applying our method has on subjective system aspects and user experience in comparison to established baselines: Learning an integrated model of latent factors and additional (user-generated) information such as tags led to significantly better scores in a majority of comparisons, emphasizing the value of *TagMF* for implementing interactive RS.

In future work, we plan to exploit the integration of user-generated tags, other content-related information (e.g. metadata on genres or keywords extracted from social media) and contextual attributes that likely affect the consumption experience (such as the current season when recommending e.g. Christmas movies) into MF more extensively. The effects of using different kinds of data as well as of enriching other recommendation methods such as deep learning with additional data still need to be investigated. In doing so, we also aim at improving current as well as developing novel application possibilities for *TagMF*. For instance, one can think of more advanced interaction mechanisms as well as improved (and possibly visually-enhanced) explanations. In line with this, we are interested in conducting further empirical user studies focusing, among others, on our tag-based explanations: As of now, we have answered RQ1d by describing in Section 4.4 a way additional content information can be used in model-based CF for explaining an existing preference profile. The derived user-tag relations should by construction describe the latent part of the user profile in an adequate manner. This is supported by the qualitative inspection we performed on the latent factor space, the successful implementation of the tag cloud in our prototype system, and the observation of participants' behavior. However, as the two present studies had a different focus, a more profound validation of this application possibility is left for future work. Moreover, although *TagMF* can be considered easily usable with other items than movies due to the domain independence of the underlying CF and the small additional requirements of our method, we especially want to evaluate it when applied to other domains.

References

[1] X. Su, T. M. Khoshgoftaar, A survey of collaborative filtering techniques, Advances in Artificial Intelligence 2009 (2009) 4:1–4:19. doi:10.1155/2009/421425.

[2] F. Ricci, L. Rokach, B. Shapira, Recommender Systems Handbook, Springer US, Boston, MA, USA, 2015, Ch. Recommender Systems: Introduction and Challenges, pp. 1–34. doi:10.1007/978-1-4899-7637-6_1.

[3] G. Jawaheer, P. Weller, P. Kostkova, Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback, ACM Transactions on Interactive Intelligent Systems 4 (2) (2014) 8:1–8:26. doi:10.1145/2512208.

[4] A. Gunawardana, G. Shani, Recommender Systems Handbook, Springer US, Boston, MA, USA, 2015, Ch. Evaluating Recommender Systems, pp. 265–308. doi:10.1007/978-1-4899-7637-6_8.

[5] B. Xiao, I. Benbasat, E-commerce product recommendation agents: Use, characteristics, and impact, MIS Quarterly 31 (1) (2007) 137–209.

[6] J. A. Konstan, J. Riedl, Recommender systems: From algorithms to user experience, User Modeling and User-Adapted Interaction 22 (1-2) (2012) 101–123. doi:10.1007/s11257-011-9112-x.

[7] P. Pu, L. Chen, R. Hu, Evaluating recommender systems from the user's perspective: Survey of the state of the art, User Modeling and User-Adapted Interaction 22 (4-5) (2012) 317–355. doi:10.1007/s11257-011-9115-7.

[8] M. Ge, C. Delgado-Battenfeld, D. Jannach, Beyond accuracy: Evaluating recommender systems by coverage and serendipity, in: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10), ACM, New York, NY, USA, 2010, pp. 257–260. doi:10.1145/1864708.1864761.

[9] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), ACM, New York, NY, USA, 2011, pp. 109–116. doi:10.1145/2043932.2043955.

[10] M. Jugovac, D. Jannach, Interacting with recommenders – Overview and research directions, ACM Transactions on Interactive Intelligent Systems 7 (3) (2017) 10:1–10:46. doi:10.1145/3001837.

[11] G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering, IEEE Internet Computing 7 (1) (2003) 76–80. doi:10.1109/MIC.2003.1167344.

[12] J. Bennett, S. Lanning, The Netflix prize, in: Proceedings of the KDD Cup and Workshop, ACM, New York, NY, USA, 2007, pp. 3–6.

[13] E. Pariser, The Filter Bubble: What the Internet is Hiding From You, Penguin Press, 2011.

[14] S. Nagulendra, J. Vassileva, Understanding and controlling the filter bubble through interactive visualization: A user study, in: Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT '14), ACM, 2014, pp. 107–115. doi:10.1145/2631775.2631811.

[15] N. Tintarev, J. Masthoff, Recommender Systems Handbook, Springer US, Boston, MA, USA, 2015, Ch. Explaining Recommendations: Design and Evaluation, pp. 353–382. doi:10.1007/978-1-4899-7637-6_10.

[16] B. P. Knijnenburg, M. C. Willemsen, Recommender Systems Handbook, Springer US, Boston, MA, USA, 2015, Ch. Evaluating Recommender Systems with User Experiments, pp. 309–352. doi:10.1007/978-1-4899-7637-6_9.

[17] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, User Modeling and User-Adapted Interaction 22 (4-5) (2012) 441–504. doi:10.1007/s11257-011-9118-4.

[18] J. Vig, S. Sen, J. Riedl, Navigating the tag genome, in: Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11), ACM, New York, NY, USA, 2011, pp. 93–102. doi:10.1145/1943403.1943418.

[19] S. Sen, J. Vig, J. Riedl, Tagommenders: Connecting users to items through tags, in: Proceedings of the 18th International Conference on World Wide Web (WWW '09), ACM, New York, NY, USA, 2009, pp. 671–680. doi:10.1145/1526709.1526800.

[20] Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, X. He, Document recommendation in social tagging services, in: Proceedings of the 19th International Conference on World Wide Web (WWW '10), ACM, New York, NY, USA, 2010, pp. 391–400. doi:10.1145/1772690.1772731.

[21] L. Chen, P. Pu, Critiquing-based recommenders: Survey and emerging trends, User Modeling and User-Adapted Interaction 22 (1-2) (2012) 125–150. doi:10.1007/s11257-011-9108-6.

[22] S. Bostandjiev, J. O'Donovan, T. Höllerer, TasteWeights: A visual interactive hybrid recommender system, in: Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12), ACM, New York, NY, USA, 2012, pp. 35–42. doi:10.1145/2365952.2365964.

[23] D. Parra, P. Brusilovsky, C. Trattner, See what you want to see: Visual user-driven approach for hybrid recommendation, in: Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14), ACM, New York, NY, USA, 2014, pp. 235–240. doi:10.1145/2557500.2557542.

[24] Y. Koren, R. M. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, IEEE Computer 42 (8) (2009) 30–37. doi:10.1109/MC.2009.263.

[25] Y. Koren, R. M. Bell, Recommender Systems Handbook, Springer US, Boston, MA, USA, 2015, Ch. Advances in Collaborative Filtering, pp. 77–118. doi:10.1007/978-1-4899-7637-6_3.

[26] T. Donkers, B. Loepp, J. Ziegler, Towards understanding latent factors and user profiles by enhancing matrix factorization with tags, in: Poster Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16), 2016.

[27] A. Karatzoglou, X. Amatriain, L. Baltrunas, N. Oliver, Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering, in: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10), ACM, New York, NY, USA, 2010, pp. 79–86. doi:10.1145/1864708.1864727.

[28] P. Forbes, M. Zhu, Content-boosted matrix factorization for recommender systems: Experiments with recipe recommendation, in: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), ACM, New York, NY, USA, 2011, pp. 261–264. doi:10.1145/2043932.2043979.

[29] J. McAuley, J. Leskovec, Hidden factors and hidden topics: Understanding rating dimensions with review text, in: Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13), ACM, New York, NY, USA, 2013, pp. 165–172. doi:10.1145/2507157.2507163.

[30] Y. Shi, M. Larson, A. Hanjalic, Mining contextual movie similarity with matrix factorization for context-aware recommendation, ACM Transactions on Intelligent Systems and Technology 4 (1) (2013) 16:1–16:19. doi:10.1145/2414425.2414441.

[31] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, C. Wang, Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS), in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14), ACM, New York, NY, USA, 2014, pp. 193–202. doi:10.1145/2623330.2623758.

[32] I. Fernández-Tobías, I. Cantador, Exploiting social tags in matrix factorization models for cross-domain collaborative filtering, in: Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems (CBRecSys '14), 2014, pp. 34–41.

[33] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, Explicit factor models for explainable recommendation based on phrase-level sentiment analysis, in: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14), ACM, New York, NY, USA, 2014, pp. 83–92. doi:10.1145/2600428.2609579.

[34] A. Almahairi, K. Kastner, K. Cho, A. Courville, Learning distributed representations from reviews for collaborative filtering, in: Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), ACM, New York, NY, USA, 2015, pp. 147–154. doi:10.1145/2792838.2800192.

[35] J. Nguyen, M. Zhu, Content-boosted matrix factorization techniques for recommender systems, Statistical Analysis and Data Mining 6 (4) (2013) 286–301. doi:10.1002/sam.11184.

[36] B. Muthén, A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, Psychometrika 49 (1) (1984) 115–132. doi:10.1007/BF02294210.

[37] E. I. Sparling, S. Sen, Rating: How difficult is it?, in: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), ACM, New York, NY, USA, 2011, pp. 149–156. doi:10.1145/2043932.2043961.

[38] P. Cremonesi, F. Garzotto, R. Turrin, User effort vs. accuracy in rating-based elicitation, in: Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12), ACM, New York, NY, USA, 2012, pp. 27–34. doi:10.1145/2365952.2365963.

[39] D. Parra, X. Amatriain, Walk the talk: Analyzing the relation between implicit and explicit feedback for preference elicitation, in: Proceedings of the 19th International Conference on User Modeling, Adaptation and Personalization (UMAP '11), Springer, Berlin, Germany, 2011, pp. 255–268. doi:10.1007/978-3-642-22362-4_22.

[40] C. He, D. Parra, K. Verbert, Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities, Expert Systems with Applications 56 (1) (2016) 9–27. doi:10.1016/j.eswa.2016.02.013.

[41] B. Loepp, C.-M. Barbu, J. Ziegler, Interactive recommending: Framework, state of research and future challenges, in: Proceedings of the 1st Workshop on Engineering Computer-Human Interaction in Recommender Systems (EnCHIReS '16), 2016, pp. 3–13.

[42] M. Mandl, A. Felfernig, Improving the performance of unit critiquing, in: Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP '12), Springer, Berlin, Germany, 2012, pp. 176–187. doi:10.1007/978-3-642-31454-4_15.

[43] B. Loepp, K. Herrmanny, J. Ziegler, Blended recommending: Integrating interactive information filtering and algorithmic recommender techniques, in: Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems (CHI '15), ACM, New York, NY, USA, 2015, pp. 975–984. doi:10.1145/2702123.2702496.

[44] C. di Sciascio, V. Sabol, E. E. Veas, Rank as you go: User-driven exploration of search results, in: Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16), ACM, New York, NY, USA, 2016, pp. 118–129. doi:10.1145/2856767.2856797.

[45] I. Andjelkovic, D. Parra, J. O'Donovan, Moodplay: Interactive mood-based music discovery and recommendation, in: Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '16), ACM, New York, NY, USA, 2016, pp. 275–279. doi:10.1145/2930238.2930280.

[46] K. Verbert, D. Parra, P. Brusilovsky, Agents vs. users: Visual recommendation of research talks with multiple dimension of relevance, ACM Transaction on Interactive Intelligent Systems 6 (2) (2016) 11:1–11:42. doi:10.1145/2946794.

[47] T. T. Nguyen, J. Riedl, Predicting users' preference from tag relevance, in: Proceedings of the 21st International Conference on User Modeling, Adaptation and Personalization (UMAP '13), Springer, Berlin, Germany, 2013, pp. 274–280. doi:10.1007/978-3-642-38844-6_23.

[48] G. E. Forsythe, M. A. Malcolm, C. B. Moler, Computer Methods for Mathematical Computations, Prentice Hall, Englewood Cliffs, NJ, USA, 1977, Ch. Least Squares and the Singular Value Decomposition.

[49] S. Rendle, L. Schmidt-Thieme, Online-updating regularized kernel matrix factorization models for large-scale recommender systems, in: Proceedings of the 2nd ACM Conference on Recommender Systems (RecSys '08), ACM, New York, NY, USA, 2008, pp. 251–258. doi:10.1145/1454008.1454047.

[50] H. Wang, N. Wang, D.-Y. Yeung, Collaborative deep learning for recommender systems, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15), ACM, New York, NY, USA, 2015, pp. 1235–1244. doi:10.1145/2783258.2783273.

[51] T. V. Nguyen, A. Karatzoglou, L. Baltrunas, Gaussian process factorization machines for context-aware recommendations, in: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14), ACM, New York, NY, USA, 2014, pp. 63–72. doi:10.1145/2600428.2609623.

[52] K. Zhou, S.-H. Yang, H. Zha, Functional matrix factorizations for cold-start recommendation, in: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11), ACM, New York, NY, USA, 2011, pp. 315–324. doi:10.1145/2009916.2009961.

[53] R. Karimi, C. Freudenthaler, A. Nanopoulos, L. Schmidt-Thieme, Exploiting the characteristics of matrix factorization for active learning in recommender systems, in: Proceedings of the 6th ACM Conference on Recommender Systems (RecSys '12), ACM, New York, NY, USA, 2012, pp. 317–320. doi:10.1145/2365952.2366031.

[54] X. Zhao, W. Zhang, J. Wang, Interactive collaborative filtering, in: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM '13), ACM, New York, NY, USA, 2013, pp. 1411–1420. doi:10.1145/2505515.2505690.

[55] M. Elahi, F. Ricci, N. Rubens, A survey of active learning in collaborative filtering recommender systems, Computer Science Review 20 (2016) 29–50. doi:10.1016/j.cosrev.2016.05.002.

[56] B. Loepp, T. Hussein, J. Ziegler, Choice-based preference elicitation for collaborative filtering recommender systems, in: Proceedings of the 32nd ACM Conference on Human Factors in Computing Systems (CHI '14), ACM, New York, NY, USA, 2014, pp. 3085–3094. doi:10.1145/2556288.2557069.

[57] M. P. Graus, M. C. Willemsen, Improving the user experience during cold start through choice-based preference elicitation, in: Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), ACM, New York, NY, USA, 2015, pp. 273–276. doi:10.1145/2792838.2799681.

[58] M. C. Willemsen, M. P. Graus, B. P. Knijnenburg, Understanding the role of latent feature diversification on choice difficulty and satisfaction, User Modeling and User-Adapted Interaction 26 (4) (2016) 347–389. doi:10.1007/s11257-016-9178-6.

[59] E. Gansner, Y. Hu, S. Kobourov, C. Volinsky, Putting recommendations on the map: Visualizing clusters and relations, in: Proceedings of the 3rd ACM Conference on Recommender Systems (RecSys '09), ACM, New York, NY, USA, 2009, pp. 345–348. doi:10.1145/1639714.1639784.

[60] J. Kunkel, B. Loepp, J. Ziegler, A 3D item space visualization for presenting and manipulating user preferences in collaborative filtering, in: Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI '17), ACM, New York, NY, USA, 2017, pp. 3–15. doi:10.1145/3025171.3025189.

[61] B. Németh, G. Takács, I. Pilászy, D. Tikk, Visualization of movie features in collaborative filtering, in: Proceedings of the 12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques (SoMeT '13), IEEE, Washington, DC, USA, 2013, pp. 229–233. doi:10.1109/SoMeT.2013.6645674.

[62] M. Rossetti, F. Stella, M. Zanker, Towards explaining latent factors with topic models in collaborative recommender systems, in: Proceedings of the 24th International Workshop on Database and Expert Systems Applications (DEXA '13), 2013, pp. 162–167. doi:10.1109/DEXA.2013.26.

[63] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, J. A. Konstan, User perception of differences in recommender algorithms, in: Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14), ACM, New York, NY, USA, 2014, pp. 161–168. doi:10.1145/2645710.2645737.

[64] D. Bollen, B. P. Knijnenburg, M. C. Willemsen, M. P. Graus, Understanding choice overload in recommender systems, in: Proceedings of the 4th ACM Conference on Recommender Systems (RecSys '10), ACM, New York, NY, USA, 2010, pp. 63–70. doi:10.1145/1864708.1864724.

[65] T. Donkers, B. Loepp, J. Ziegler, Merging latent factors and tags to increase interactive control of recommendations, in: Poster Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), 2015.

[66] E. H. Moore, On the reciprocal of the general algebraic matrix, Bulletin of the American Mathematical Society 26 (1920) 394–395. doi:10.1090/S0002-9904-1920-03322-7.

[67] R. Penrose, A generalized inverse for matrices, Mathematical Proceedings of the Cambridge Philosophical Society 51 (3) (1955) 406–413. doi:10.1017/S0305004100030401.

[68] G. Takács, I. Pilászy, B. Németh, D. Tikk, Scalable collaborative filtering approaches for large recommender systems, Journal of Machine Learning Research 10 (2009) 623–656.

[69] A. Said, A. Bellogín, RiVal: A toolkit to foster reproducibility in recommender system evaluation, in: Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14), ACM, New York, NY, USA, 2014, pp. 371–372. doi:10.1145/2645710.2645712.

[70] T. Donkers, B. Loepp, J. Ziegler, Tag-enhanced collaborative filtering for increasing transparency and interactive control, in: Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '16), ACM, New York, NY, USA, 2016, pp. 169–173. doi:10.1145/2930238.2930287.

[71] J. Vig, S. Sen, J. Riedl, Computing the tag genome, Tech. rep., University of Minnesota (2010).

[72] B. P. Knijnenburg, M. C. Willemsen, A. Kobsa, A pragmatic procedure to support the user-centric evaluation of recommender systems, in: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), ACM, New York, NY, USA, 2011, pp. 321–324. doi:10.1145/2043932.2043993.

[73] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11), ACM, New York, NY, USA, 2011, pp. 157–164. doi:10.1145/2043932.2043962.

[74] J. Brooke, SUS – A quick and dirty usability scale, in: Usability Evaluation in Industry, Taylor & Francis, London, UK, 1996, pp. 189–194.

[75] B. Laugwitz, T. Held, M. Schrepp, Construction and evaluation of a user experience questionnaire, in: Proceedings of the 4th Symposium of the Austrian HCI and Usability Engineering Group (USAB '08), Springer, Berlin, Germany, 2008, pp. 63–76. doi:10.1007/978-3-540-89350-9_6.

[76] B. P. Knijnenburg, M. C. Willemsen, The effect of preference elicitation methods on the user experience of a recommender system, in: Extended Abstracts of the 28th ACM Conference on Human Factors in Computing Systems (CHI '10), ACM, New York, NY, USA, 2010, pp. 3457–3462. doi:10.1145/1753846.1754001.

[77] T. T. Nguyen, D. Kluver, T.-Y. Wang, P.-M. Hui, M. D. Ekstrand, M. C. Willemsen, J. Riedl, Rating support interfaces to improve user experience and recommender accuracy, in: Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13), ACM, New York, NY, USA, 2013, pp. 149–156. doi:10.1145/2507157.2507188.

[78] M. D. Ekstrand, D. Kluver, F. M. Harper, J. A. Konstan, Letting users choose recommender algorithms: An experimental study, in: Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15), ACM, New York, NY, USA, 2015, pp. 11–18. doi:10.1145/2792838.2800195.

[79] J. Vig, S. Sen, J. Riedl, The tag genome: Encoding community knowledge to support novel interaction, ACM Transactions on Interactive Intelligent Systems 2 (3) (2012) 13:1–13:44. doi:10.1145/2362394.2362395.

[80] A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan, J. Riedl, Getting to know you: Learning new user preferences in recommender systems, in: Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI '02), ACM, New York, NY, USA, 2002, pp. 127–134. doi:10.1145/502716.502737.

[81] L. Iaquinta, M. d. Gemmis, P. Lops, G. Semeraro, M. Filannino, P. Molino, Introducing serendipity in a content-based recommender system, in: Proceedings of the 8th International Conference on Hybrid Intelligent Systems (HIS '08), IEEE, Washington, DC, USA, 2008, pp. 168–173. doi:10.1109/HIS.2008.25.

[82] J. Vig, S. Sen, J. Riedl, Tagsplanations: Explaining recommendations using tags, in: Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09), ACM, New York, NY, USA, 2009, pp. 47–56. doi:10.1145/1502650.1502661.