Integrating the ACT-R Framework with Collaborative Filtering for Explainable Sequential Music Recommendation

Marta Moscati
marta.moscati@jku.at
Institute of Computational Perception, Johannes Kepler University Linz
Linz, Austria

Christian Wallmann
ch.wallmann@welser.com
Welser Profile GmbH
Gresten, Austria

Markus Reiter-Haas
reiter-haas@tugraz.at
Graz University of Technology
Graz, Austria

Dominik Kowald
dkowald@know-center.at
Know-Center GmbH and Graz University of Technology
Graz, Austria

Elisabeth Lex
elisabeth.lex@tugraz.at
Graz University of Technology
Graz, Austria

Markus Schedl
markus.schedl@jku.at
Institute of Computational Perception, Johannes Kepler University Linz and Human-centered AI Group, AI Lab, Linz Institute of Technology
Linz, Austria

ABSTRACT

Music listening sessions often consist of sequences including repeating tracks. Modeling such relistening behavior with models of human memory has been proven effective in predicting the next track of a session. However, these models intrinsically lack the capability of recommending novel tracks that the target user has not listened to in the past. Collaborative filtering strategies, on the contrary, provide novel recommendations by leveraging past collective behaviors but are often limited in their ability to provide explanations. To narrow this gap, we propose four hybrid algorithms that integrate collaborative filtering with the cognitive architecture ACT-R. We compare their performance in terms of accuracy, novelty, diversity, and popularity bias, to baselines of different types, including pure ACT-R, kNN-based, and neural-networks-based approaches. We show that the proposed algorithms are able to achieve the best performances in terms of novelty and diversity, and simultaneously achieve a higher accuracy of recommendation with respect to pure ACT-R models. Furthermore, we illustrate how the proposed models can provide explainable recommendations.

KEYWORDS

Adaptive Control Thought-Rational (ACT-R), Collaborative Filtering, Sequential Recommendation, Music Recommender Systems, Psychology-Informed Recommender Systems, Explainability

ACM Reference Format:
Marta Moscati, Christian Wallmann, Markus Reiter-Haas, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2023. Integrating the ACT-R Framework with Collaborative Filtering for Explainable Sequential Music Recommendation. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3604915.3608838

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3608838

1 INTRODUCTION

Music is often consumed sequentially. Therefore, music recommendation [46, 48] is often formulated as a session completion task: tracks should be recommended to a user according to their interactions in the recent past, i.e., those within the current session. The most effective recommender systems (RSs) for sequential recommendation are based on collaborative filtering (CF) [11, 14, 32, 33]. These algorithms provide recommendations according to past collective user behavior. Although effective, the recommendations provided by CF algorithms are often hard to justify, either due to the model architecture or the complexity of the data they base their recommendations on. Another major distinguishing characteristic of music RSs compared to general RSs is that music listeners often listen to tracks they already listened to in the past [7, 42, 48]. This observation served as a basis for translating cognitive architectures, i.e., models of the structure of human mind, to the domain of RSs, and evaluate their effectiveness in predicting users' relistening behaviors. In particular, the memory module of the *Adaptive Control of Thought—Rational* (ACT-R) cognitive architecture [6, 44] has been proven effective in predicting which tracks the user will relisten to, based on the tracks listened to in the past [40]. However, despite their effectiveness in modeling user's relistening behavior, leveraging these models based on ACT-R for sequential music recommendation does not allow recommending *novel* tracks, i.e., tracks the target user has never interacted with before. To compensate for these shortcomings, we design four algorithms that integrate ACT-R with CF. Since each component of the memory module of ACT-R is designed to model a different aspect of human memory, the recommendations provided by the proposed algorithms are explainable. We measure the performance of the proposed RSs in terms of accuracy, novelty, diversity, and popularity bias of the recommended tracks since these are all aspects that affect the user's satisfaction with the system [16, 48]. Additionally, we show how the explainability of the proposed algorithms can be advantageous in a

multistakeholder RS [1], concerning end users, platform providers, and content producers.

In summary, this work provides the following contributions to the RS domain: (1) We propose four algorithms that integrate various components of the cognitive architecture ACT-R with CF for sequential recommendation. (2) We provide an extensive analysis of the performance of these algorithms by performing experiments on the LFM-2b dataset [45] of Last.fm listening logs. We compare the performance of the algorithms with well-established baselines, including algorithms that solely rely on the cognitive architecture ACT-R, on k-nearest-neighbors (kNN), and on deep neural networks (DNNs). Our experiments show that the proposed algorithms increase the novelty and diversity of recommendations compared to the baselines. Moreover, we find that the hybrid approaches outperform pure ACT-R models in terms of accuracy. (3) We exemplify how the proposed algorithms can be used to explain music recommendations.

2 BACKGROUND AND RELATED WORK

In the following, we briefly present work on sequential recommendation and on RSs based on cognitive architectures, thereby introducing the fundamentals for the proposed algorithms.

2.1 Sequential Recommendation

Some of the most successful sequential RSs leverage the similarity of the initial segment of the session to be completed to other sessions. Extensions of these algorithms also introduce *temporal reweighting*, i.e. they consider factors that model the position and recency of the interactions with the items [11, 31, 34, 41]. For instance, Ludewig et al. [34] reweight the recommendation score as follows: if an item i appeared at position t_i, its relevance as a recommendation for position t_{ref} is weighted by a factor given by $w_i = (t_{\text{ref}} - t_i)^{-d}$, where t_{ref} stands for the timestamp of the next track, i.e., the one the algorithm is aiming to predict. Some effective approaches use DNN architectures for sequential recommendation [8, 13, 17, 27, 30, 49, 50]. Finally, other works model the sessions by representing them as graphs, and leverage graph neural networks [5, 12, 38, 50–52].

2.2 Music Recommender Systems

Sequential RSs are particularly relevant in the context of music recommendation [47, 48] since they address tasks such as next-track recommendation or automatic playlist continuation. For an overview of the approaches used for sequential music recommendation we refer the reader to Quadrana et al. [39]. Additionally, since providing explanations for the recommendations can positively impact the users' trust and engagement, the interest in addressing explainability in the context of music RSs has been increasing in the last years; for an overview of the topic we refer the reader to Afchar et al. [4].

2.3 Cognition-inspired Recommender Systems

Cognition-inspired RSs use models from the domain of cognitive psychology to create RSs, often using theories of human memory [26, 28]. Our work focuses on ACT-R [6, 44]. Several studies leveraged the memory module of ACT-R for tasks such as hashtag recommendation [19], item recommendation in social tagging systems [21], next genre prediction [25], next artist prediction [18], job recommendation [20, 22], or predicting mobile app usages [53]. In particular, Reiter-Haas et al. [40] use ACT-R's memory module for completing music streaming sessions. The components of the module are described in the remainder of this section.

Base-Level Learning (BLL): The BLL component captures the tendency of human memory of favoring instances that occurred frequently and recently in the past. Similar to Reiter-Haas et al. [40], given the timestamp t_{ref} of the next track in the session, i.e., the one the algorithm is aiming to predict, and given an item i, we define its BLL activation as $B_i = \sum_{j=1}^{n} (t_{\text{ref}} - t_{ij})^{-d}$. The sum extends to all the n past interactions of the user with item i, and t_{ij} stands for the timestamp of the j^{th} interaction with item i.

Spreading (S): The spreading component favors items that occur frequently in the current *context*. In agreement with how context is defined within the ACT-R cognitive architecture, Reiter-Haas et al. [40] define the context as the last item the user interacted with. This component hence tends to favor items that the user often interacted with in sessions that contain the most recent item in the sequence. The corresponding activation is given by $S_i = \frac{P(i \in C_k)}{P(i)}$ [10, 40], where item k is the last item of the sequence, and $P(i)$ and $P(i \in C_k)$ stand for the probabilities that track i appears in any session, and in a session containing item k, respectively.

Partial Matching (PM): The PM component [40] aims at favoring items that are *similar* to the context item k, i.e., the last item the user interacted with. The corresponding activation is given by $P_i = \text{sim}(i, k)$, where $\text{sim}(i, k)$ represents the similarity between item i and the context item k. Assuming an item i to be represented by a feature vector $\mathbf{f}_i$, the similarity $\text{sim}(i, k)$ between i and k is defined as the scalar product of the corresponding feature vectors, $\text{sim}(i, k) = \mathbf{f}_i \cdot \mathbf{f}_k$.

Valuation (V): The valuation component [15, 40] aims at measuring the *value* attributed by a user to an item. The corresponding activation for an item i with which the user interacted n times is defined iteratively as $V_i(n) = V_i(n-1) + \alpha\,(R_i(n) - V_i(n-1))$, where $R_i(n)$ is the reward assigned to item i for the n^{th} interaction. The starting valuation is set to $V_i(0) = 0$ for all tracks, and the learning rate α is considered as a hyperparameter. In the context of sequential music recommendation [40], the reward $R_i(j)$ is typically either binary, i.e., $R_i(j) = 1 \forall j \in [1, \ldots, n]$, or given by the duration of the j^{th} interaction with respect to the total track length.

Noise (N): The noise component models aspects of randomness in the user's behavior. The corresponding activation is given by $\epsilon_i = \text{rng}()$, where $\text{rng}()$ is a random number generator.

Reiter-Haas et al. [40] show that ACT-R-based approaches outperform baselines such as algorithms selecting the most recent track, in terms of accuracy of predictions. Compared to their work, we integrate ACT-R and CF, extend the analysis to beyond-accuracy metrics, and provide a comparison with more recent baselines. Finally, we also leverage ACT-R for explaining the recommendations.

3 METHODS

To integrate ACT-R and CF for sequential music recommendation, we propose the following hybrid algorithms.

`Social ACT-R` Kowald et al. [19] propose an algorithm for hashtag recommendation that combines the ACT-R activations of the target user's past hashtags with the ACT-R activations of the target user's followees.

We adapt this strategy to the music domain. In order to include the listening behavior of other users, we first define the target user's "followees" as the set of k users that are most similar to the target user. The similarity $\text{sim}_{\texttt{ACT-R}}(u, j)$ between the target user u and another user j is computed as cosine similarity between the vector representing their listening events, i.e., their interactions with tracks, reweighted with the ACT-R activations. The value of the social component (SC) assigned to track i for session u is then defined as a similarity-weighted average of the ACT-R activations of the k followees, $SC_i = \sum_{j \leq k} \texttt{ACT-R}(j, i) \cdot \text{sim}_{\texttt{ACT-R}}(u, j)$. The SC and the target user's ACT-R activation of the track are normalized by applying softmax over all tracks and added up to obtain the final recommendation score.

ACT-R + BPR This model extends ACT-R with a component that favors tracks that have a similar interaction history to the one of the context track, i.e., the last track the user listened to. For this purpose, we pretrain a matrix factorization RS with Bayesian personalized ranking (BPR) [43]. Each track is mapped to their BPR embedding v_i. We then compute the similarity $\text{sim}_{\texttt{BPR}}(i, j)$ between two tracks i, j as cosine similarity between their BPR embeddings v_i, v_j. The recommendation score of a track i is obtained by adding up the softmax-normalized BPR similarity $\text{sim}_{\texttt{BPR}}(i, k)$ with the context item k and the softmax-normalized target user's ACT-R activation of i. In addition, we consider a version of this model in which only the similarity $\text{sim}_{\texttt{BPR}}(i, k)$ between the BPR embeddings is considered when computing the recommendation score. This model is referred to as Item BPR.

Weighted MultVAE: We integrate ACT-R with MultVAE [29], since this model allows providing recommendations to users that are not in the train set, and since it provides accurate recommendations in several domains, including music [9, 36]. We pretrain and optimize an instance of MultVAE. We then reweight the components of the vector representing the listening events of the target user u either with the ACT-R activations or with the temporal reweighting factor w_i (see Section 2.1), converting it to a vector of ratings. We feed this vector to the pretrained MultVAE and perform a forward pass of MultVAE to select the tracks to recommend.

Weighted UserkNN: Similar to Weighted MultVAE, we first train an instance of MultVAE and then perform a forward pass on the temporally reweighed vector representing the listening events of the target user, extracting the latent representations l_u of the target user u encoded by MultVAE. We encode the binarized[1] profile of the other users in the dataset and select the k users with latent representations having the largest cosine similarity $\text{sim}_{\texttt{MultVAE}}(u, j)$ to the latent representation l_u of the target user. We take the weighted average of the binarized profiles of the k nearest users, using the similarity of the latent representations as weights for the weighted average, as recommendation score. The score of track i is therefore given by $\sum_{j \leq k} r(j, i) \cdot \text{sim}_{\texttt{MultVAE}}(u, j)$, where $r(j, i)$ represents the binarized interaction of user j with item i.

[1] In agreement with the reweighing proposed by Ludewig et al. [34], we do not apply temporal reweighing to the profile of the non-target users.

4 EXPERIMENTAL SETUP

In this section, we describe the setup for our experiments, i.e., the baseline models, the evaluation metrics, the dataset, as well as the training and hyperparameter selection.

4.1 Baselines and underlying models

We compare the performance of the approaches introduced in Section 3 to those of two models effective in the task of sequential recommendation – GRU4Rec [49] and temporal UserkNN [32, 33] – and two models effective in predicting relistening behavior – MostRecent [40] and ACT-R [40].

GRU4Rec: This algorithm makes use of recurrent neural networks for sequential recommendation [49]. We take it as DNN-based baseline since it is among the DNN approaches achieving high accuracy, large dataset coverage, and low popularity bias, simultaneously [33].

Temporal UserkNN: Models including a temporal reweighting (see Section 2.1) are competitive with DNN-based approaches in terms of accuracy [33, 34]. In including this class of models as baselines, we reweight the vectors representing the listening events of the target user, as well as those representing the other users, as described in Section 2.1. We then compute the cosine similarity of the resulting vectors. The reweighted interactions of the k nearest users are averaged according to the similarity to the target user and used as recommendation scores.

MostRecent: This algorithm recommends the most recent tracks in the sequence, and has been proven effective in predicting users' relistening behavior [40], especially in accurately predicting the next track in the session (see discussion of Next-HR in Section 4.2).

ACT-R: This model corresponds to the one used by Reiter-Haas et al. in [40] for modeling the users' music relistening behavior; we refer the reader to Section 2.3 for the description of the individual components.[2]

4.2 Evaluation metrics

The performance of the algorithms is evaluated on the task of rolling session completion. Similar to Reiter-Haas et al. [40], for each target user we shift a sliding window of one week with a hop size of one listening eventand define sessions as sequences of listening events without gaps of more than 30 minutes between consecutive tracks. Given a target user's session of N tracks and the target user's listening events of the previous seven days, we assume a session segment of length $l < N$ to be known and predict the remaining $N - l$ tracks in the session. For each session, we consider all possible initial segment lengths, $l = 1, \ldots, N - 1$.

Accuracy: We include two metrics for the accuracy of recommendations. Since ACT-R can only recommend items that the user already listened to, if the number of past interactions is less than the number of tracks in the remainder of the sessions, i.e., those to provide recommendations for, the algorithm will not be able to provide recommendations for the full session. This results in a higher precision and a lower recall. To mitigate this effect, we

[2] Similar to Reiter-Haas et al. [40], we normalize each component by applying softmax over all tracks and add up the results to obtain the ACT-R activation of a track. For all ACT-R-based models, preliminary experiments showed that including ϵ and PM based on different versions of the features provided by Spotify reduces the performance of the algorithms. We, therefore, omit them. Based on preliminary experiments, we also set $d = 0.5$.

combine the precision and recall of recommendations into the F_1 score. Similar to Reiter-Haas et al. [40], we also evaluate the *next hit rate* (Next-HR), i.e., the ability of the algorithm to correctly predict the next track of the session.

Novelty: The novelty of recommendations is measured as the fraction of recommended tracks that have not been listened to by the target user. In addition, to evaluate the quality of novel recommendations, we also report the precision of novel recommendations, P-Novelty.

Diversity: The diversity of the recommendations is measured with respect to the genres.[3] Since a higher diversity should indicate that the recommended tracks belong to different genres, we define diversity as the Shannon entropy of the distribution of genres over the recommended tracks.

Popularity bias: To evaluate the tendency of the algorithms to overrepresent popular tracks compared to the ones in the user's past listening events, we compute the Jensen-Shannon divergence between the popularity distribution of tracks in the user's past listening events, and over the recommended tracks [23]. A high popularity bias thus indicates that recommended tracks are more popular than those already listened to by the target user. [4]

4.3 Dataset, training, and evaluation

Similar to Melchiorre et al. [35], we conduct our experiments on the extract of the large LFM-2b dataset [45][5] corresponding to the last month (20/02/2020 - 19/03/2020) and remove users that listened to more tracks than the 99$^{\text{th}}$ percentile of all users. We apply 10-core filtering to users and items and split each user's listening events temporally in a 60% train, 20% validation, and a 20% test set. The resulting dataset consists of 2 889 028 listening events, 12 679 users and 101 837 items. The 60% train set is used to determine the similarity for approaches relying on kNN, and for training and selecting the best hyperparameters of GRU4Rec, BPR, and MultVAE. For GRU4Rec, the most recent 20% interactions of each user in the 60% train set are used for selecting the best hyperparameter configuration on a grid space based on that reported by Ludewig et al. [33, 34]. The optimization of the BPR and MultVAE instances required by ACT-R + BPR, Weighted MultVAE, and Weighted MultVAE-UserkNN is performed previously to the optimization of the algorithms that rely on them, and therefore on a separate set: the 60% train set is converted to a binarized version (i.e., $b_{ui} = 1 \iff u$ listened to i at least once) and 20% randomly selected binarized interactions of each user are used for selecting thehyperparameter configuration achieving the highest NDCG@10 on a grid space based on that reported by Melchiorre et al. [36]. The 20% validation set is used to select the best configuration of kNN-, temporal, and ACT-R-based algorithms described in Sections 3 and 4.1, on a grid space based on that reported by Ludewig et al. [33, 34] and Reiter-Haas et al. [40], selecting the configuration achieving the highest F_1 score. Since the average number of session completion tasks per user is 67, providing recommendations for the sessions of all 12 679 users would result in roughly 850 000 session recommendations, i.e., roughly 850 000 test users in a standard recommendation scenario. Therefore, to reduce computational costs, we randomly sample 100 users and evaluate the algorithms' performances reported in Section 5 on the corresponding test sessions, for a total of 6 697 session completion tasks.[6]

5 PERFORMANCE COMPARISON

Table 1 reports the performance of the algorithms in terms of accuracy, novelty, diversity, and popularity bias on the 6 697 test session completions of the 100 randomly sampled users. In terms of accuracy (F_1 and Next-HR), GRU4Rec outperforms the proposed algorithms, as well as the other baselines. In terms of F_1, GRU4Rec is followed by Temporal UserkNN; this confirms the results from previous work [14, 33], showing that these two algorithms are competitive in the task of sequential recommendation. Interestingly, however, when looking at Next-HR the performance of Temporal UserkNN displays a substantial drop, and is clearly outperformed by MostRecent. This confirms the results reported by Reiter-Haas et al. [40], indicating that recommending the last track of the session is a strong baseline with respect to Next-HR. The fact that approaches based on recurrent neural networks achieve high accuracy of recommendation, and that simply recommending the last track achieves a high Next-HR, indicate that listening sessions tend to display recurring temporal patterns (in the extreme case, repetitions of single tracks). With respect to P-Novelty, two of our proposed algorithms achieve the best performances: ACT-R + BPR and Social ACT-R. Social ACT-R achieves the highest values of diversity, indicating that it is able to recommend tracks of various genres for completing a session. It is interesting to observe that both diversity and accuracy of Social ACT-R recommendations are higher compared to ACT-R: The inclusion of collaborative information in Social ACT-R hence increases diversity and F_1 simultaneously. Finally, we observe that simple temporal- or memory-based approaches, i.e., MostRecent and ACT-R, are less biased towards popular tracks. This pattern could be explained by the fact that, since they only consider the listening events of the target user, they do not rely on collaborative data, which is a common source of popularity bias [2, 3, 24, 37].

In summary, we observe that the proposed algorithms – although not outperforming the accuracy of DNN-based algorithms – achieve the highest performance in terms of beyond-accuracy metrics, such as novelty and diversity, are able to provide more accurate novel recommendations (P-Novelty), and outperform pure ACT-R models in terms of accuracy and beyond-accuracy metrics.

6 EXPLAINABLE MUSIC RECOMMENDATION

Since the proposed algorithms are based on a well-defined psychological model, their recommendations are intrinsically explainable. This is advantageous for different RS stakeholders, as we discuss in this section.

[3]The track genre is assigned based on the Last.fm tags of the track, selecting the genre with the highest tag weight. We use the list of Discogs genres, available at https://mtg.github.io/acousticbrainz-genre-dataset/data_stats/ as possible genres of a track.

[4]The popularity of a track is defined as the ratio of the total number of listening events it accounts for [3, 24]. The distributions are computed over popularity classes, each defined in terms of percentiles: after sorting tracks according to the number of listening events, popular-, mid-, and niche-tracks account for 20, 60, and 20% of all events, respectively.

[5]http://www.cp.jku.at/datasets/LFM-2b/

[6]For reproducibility purposes, we share the code, dataset, details on the dataset handling and splits, hyperparameter optimization, and pretrained instances of BPR and MultVAE required by the algorithms described in Section 3 at https://github.com/hcai-mms/actr.

Table 1: Performance of the models in the session-completion task. Models are sorted in order of descending F_1 score. All values are averaged over the 6 697 test session completions of the 100 randomly sampled users, as described in Section 4.3 New models are highlighted in blue. Best performances are highlighted in bold, second best are underlined.

	F_1	Next-HR	Novelty	P-Novelty	Diversity	PopBias
GRU4Rec	**0.142**	**0.198**	0.716	0.126	0.929	0.099
Temporal UserkNN	0.122	0.024	0.631	0.146	0.786	0.122
Item BPR	0.114	0.051	**0.846**	0.130	0.658	0.151
Weighted MultVAE	0.111	0.045	0.554	0.136	0.941	0.194
ACT-R + BPR	0.104	0.037	0.056	**0.239**	0.923	0.065
Social ACT-R	0.101	0.037	0.056	0.155	**0.945**	0.066
MostRecent	0.094	0.069	0.000	0.000	0.891	**0.060**
ACT-R	0.093	0.037	0.000	0.000	0.889	**0.060**
Weighted MultVAE UserkNN	0.064	0.010	0.831	0.064	0.833	0.176

Figure 1a shows an example of an initial segment of length $l = 6$, of a session of total length $N = 12$. The initial segment consists of five unique tracks, one of them listened to twice. Figure 1b shows the list of $N - l = 6$ tracks with the highest recommendation score according to Social ACT-R. The columns show the relative contribution of the components of Social ACT-R, also reflected as a color gradient. Each component captures a different aspect of relevance for Social ACT-R's recommendations, and can therefore be translated in a way that can easily be understood by the end user. *Current obsession* corresponds to BLL, which captures the recency (*Current*) and frequency (*obsession*) of interactions with the track. *Current vibes* corresponds to S since this component favors tracks that often occurred together with the last one. *Evergreens* corresponds to V which, with a binary reward, favors tracks that were often listened to by the target user, irrespective of when in the past. Finally, *From similar listeners* corresponds to SC, which reflects collaborative information. Figure 1 hence gives a clear indication of why each song was recommended to the user. The top-5 recommendations all belong to the target user's current session. The 6^{th} is a track that the target user never listened to, as it is evident from the vanishing ACT-R components. In this particular case, the ACT-R scores all vanish for other elements of the catalog, indicating that in the one-week window used to evaluate the ACT-R scores, only the current session is present. Therefore, in this example ACT-R alone would not allow recommending more than five tracks, while this can be achieved with Social ACT-R. For instance, the track at the top of the recommendation list (*From the Past comes the Storms*) appeared in the target user's past interactions recently (BLL), often (V and BLL), and in sessions that included the most recent track the user listened to (S). For the last track in the list, the situation is different: this track is not part of the target user's past interactions; therefore it was recommended since users with a similar listening profile (according to the ACT-R activations) also listened to it. The possibility to investigate the contributions of the different components to the final recommendation score may also provide useful information to the platform providers. In the example provided in Figure 1, for instance, we see that the ACT-R components entirely contribute to the recommendation scores at the top of the list, while SC only becomes relevant once all the tracks of the initial segment of the current session have been recommended. We attribute this to the fact that the ACT-R activations are nonvanishing for a limited set of tracks (all tracks the target user listened to in the last seven days), while SC does not vanish for a larger set of tracks (all tracks listened to by the k most similar users). Aggregating the individual ACT-R activations and SC as described in Section 3 hence results in a very peaked individual ACT-R distribution over items, and a more spread and almost negligible SC. Therefore, if a platform provider wants to favor the CF component, for instance for providing more novel recommendations, they might consider rank-based aggregation techniques, or consider assigning a higher weight to SC. The explainability of RSs that integrate ACT-R with CF can also be used to discover patterns in recommendations, which can be useful both to platform providers and content producers. To give an example, we analyze how often each of the Social ACT-R components is the *salient* one, i.e., how often the score of this component is larger than the score of the other components. Figure 2 shows the salience of each component, in percentage over all recommendations, and for specific genres. By looking at the salience over all genres, we see that taken together, the ACT-R components are salient for about 90% of all recommendations, with S being the dominant component for more than half of them. Hence, context, i.e., the last track the user listened to, often plays the largest role in selecting which track to recommend. This tendency can change when looking at specific genres. For instance, while for tracks of the genre *non-music* – often corresponding to spoken words – the salience of S is even increased, the situation is inverted for *stage and screen* – e.g., tracks that are part of movie soundtracks. For *stage and screen*, BLL is often the salient component, and together with V is the salient component in above 75% of recommendations. This indicates that for recommendations of *non-music* tracks, the last track listened to is particularly relevant, while for *stage and screen* frequency of occurrence in the past listening events is more important. This information can be further leveraged by the platform providers to weight the relative importance of each component in a genre-specific way, in order to design ACT-R- and content-based RSs, whose recommendations are tailored to genres. Finally, investigating the components' salience may help content producers to gain insight into the behavior of listeners of specific genres. For instance, *reggae* artists might observe that for this genre V is often the most salient contribution

Session Position	Listened Track
1	The Abyss
2	R.I.P. (Rest in Pain)
3	From the Past Comes the Storms
4	From the Past Comes the Storms
5	To the Wall
6	Escape the Void

(a) Initial segment of length $l = 6$.

Recommended Track	Current obsession (BLL)	Current vibes (S)	Evergreens (V)	From similar listeners (SC)
From the Past Comes the Storms	0.471	0.248	0.281	0.000
Escape to the Void	0.306	0.353	0.341	0.000
To the Wall	0.294	0.359	0.347	0.000
R.I.P. (Rest in Pain)	0.264	0.374	0.362	0.000
The Abyss	0.263	0.375	0.362	0.000
Troops of Doom	0.000	0.000	0.000	1.000

(b) Recommendations for the remaining $N - l = 6$ tracks in the session.

Figure 1: Left: Example of initial segment of length 6 of a target session of total length 12. The column "Session Position" displays the position of the track in the initial segment of the target session. Right: Heatmap of the relative contribution of the used `Social ACT-R` components to the total recommendation score of each of the 6 recommendations (remaining session length). The more intense the color, the higher the contribution.

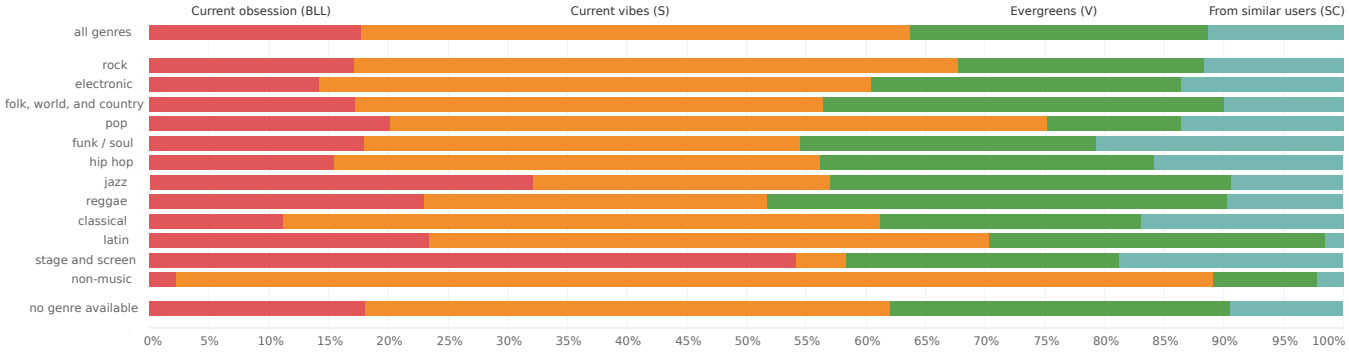

Figure 2: Component *salience* over all `Social ACT-R` recommendations and over `Social ACT-R` recommendations of a specific genre. A component is considered salient if its score is higher than the scores of the other components of the same track. Investigating the components' salience may help content producers to understand their listeners' behaviour.

and conclude that their fans have a higher tendency to relisten to the same tracks, irrespective of the last track they listened to.

7 CONCLUSION AND FUTURE WORK

In this work, we proposed four new RS algorithms that integrate the ACT-R cognitive architecture with CF for sequential music recommendation: `Social ACT-R`, `ACT-R + BPR`, Weighted `MultVAE`, and Weighted `MultVAE UserkNN`. We showed that although the proposed algorithms do not outperform the accuracy of DNN-based recommenders, they achieve the highest performance in terms of beyond-accuracy metrics. In particular, integrating CF with `ACT-R` in `Social ACT-R` achieves the highest diversity and simultaneously increases F_1 with respect to `ACT-R`. More importantly, the proposed algorithms can be used for providing explainable recommendations, which can enhance the users' engagement with the platform, provide insight to platform providers on the RSs, and to artists on the listening behaviors of their listeners. One of the limitations of this work is that it exclusively considers one perspective from cognitive psychology, i.e., that of the ACT-R model. Additionally, the definition of context for the spreading and partial matching components is given in terms of the last item of the session. While this agrees with the way context is defined in the ACT-R cognitive architecture, it would be interesting to extend the work with definitions of context that are more common in the RS community, such as location or time of the day. Moreover, we optimized the RSs for achieving the highest F_1 score. Due to the structure of the proposed algorithms, including beyond-accuracy metrics in the optimization process would allow analyzing how each component impacts the different aspects of recommendation. This could be translated to more detailed explanations and be leveraged for the design of a hybrid RS that can be tuned by each user according to their needs. We leave these extensions of our work for future research. Finally, the explainability of the algorithms proposed in this work opens up the possibility to evaluate the quality, user acceptance and understanability of the explanations by means of user studies; we leave this evaluation for future work.

ACKNOWLEDGMENTS

This research was funded in whole, or in part, by the Austrian Science Funds (FWF): P33526 and DFH-23, and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research, through grant LIT-2020-9-SEE-113.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30 (2020), 127–158.

[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proc. of ACM RecSys* (Como, Italy). 42–46.

[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-Centered Evaluation of Popularity Bias in Recommender Systems. In *Proc. of ACM UMAP* (Utrecht, Netherlands). 119–129.

[4] Darius Afchar, Alessandro B. Melchiorre, Markus Schedl, Romain Hennequin, Elena V. Epure, and Manuel Moussallam. 2022. Explainability in music recommender systems. *AI Magazine* 43, 2 (2022), 190–208.

[5] Mehrnaz Amjadi, Seyed Danial Mohseni Taheri, and Theja Tulabandhula. 2021. KATRec: Knowledge Aware aTtentive Sequential Recommendations. In *Proc. of DS*. 305–320.

[6] John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. An integrated theory of the mind. *Psychological review* 111, 4 (2004), 1036.

[7] Frederick Conrad, Jason Corey, Samantha Goldstein, Joseph Ostrow, and Michael Sadowsky. 2019. Extreme re-listening: Songs people love... and continue to love. *Psychology of Music* 47, 2 (2019), 158–172.

[8] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. 2021. Transformers4Rec: Bridging the Gap between NLP and Sequential / Session-Based Recommendation. In *Proc. of ACM RecSys* (Amsterdam, Netherlands). 143–153.

[9] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. In *Proc. of ACM RecSys*. 101–109.

[10] Danilo Fum and Andrea Stocco. 2004. Memory, Emotion, and Rationality: An ACT-R interpretation for Gambling Task results. In *Proc. of ICCM* (Pittsburgh, PA, USA). 106–111.

[11] Diksha Garg, Priyanka Gupta, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. 2019. Sequence and Time Aware Neighborhood for Session-based Recommendations: STAN. In *Proc. of ACM SIGIR* (Paris, France). 1069–1072.

[12] Tajuddeen Rabiu Gwadabe and Ying Liu. 2022. IC-GAR: item co-occurrence graph augmented session-based recommendation. *Neural Computing and Applications* 34, 10 (2022), 1–16.

[13] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel Recurrent Neural Network Architectures for Feature-rich Session-based Recommendations. In *Proc. of ACM RecSys* (Boston, MA, USA). 241–248.

[14] Dietmar Jannach and Malte Ludewig. 2017. When Recurrent Neural Networks meet the Neighborhood for Session-Based Recommendation. In *Proc. of ACM RecSys* (Como, Italy). 306–310.

[15] Ion Juvina, Othalia Larue, and Alexander Hough. 2018. Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. *Cognitive Systems Research* 48 (2018), 4–24.

[16] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems* 7, 1 (2016), 42 pages.

[17] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *Proc. of IEEE ICDM*. 197–206.

[18] Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2019. Modeling artist preferences of users with different music consumption patterns for fair music recommendations. In *LBR of ISMIR* (Delft, Netherlands).

[19] Dominik Kowald, Subhash Chandra Pujari, and Elisabeth Lex. 2017. Temporal effects on hashtag reuse in twitter: A cognitive-inspired hashtag recommendation approach. In *Proc. of ACM WWW* (Tacoma, WA USA). 1401–1410.

[20] Emanuel Lacic, Dominik Kowald, Markus Reiter-Haas, Valentin Slawicek, and Elisabeth Lex. 2017. Beyond Accuracy Optimization: On the Value of Item Embeddings for Student Job Recommendations. *CoRR* (2017).

[21] Emanuel Lacic, Dominik Kowald, Paul Christian Seitlinger, Christoph Trattner, and Denis Parra. 2014. Recommending Items in Social Tagging Systems Using Tag and Time Information. In *In Proceedings of the 1st Social Personalization Workshop co-located with the 25th ACM Conference on Hypertext and Social Media*. Association of Computing Machinery, 4–9.

[22] Emanuel Lacic, Markus Reiter-Haas, Tomislav Duricic, Valentin Slawicek, and Elisabeth Lex. 2019. Should we embed? A study on the online performance of utilizing embeddings for real-time job recommendations. In *Proc. of ACM RecSys* (Copenhagen, Denmark). 496–500.

[23] Oleg Lesota, Stefan Brandl, Matthias Wenzel, Alessandro B. Melchiorre, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2022. Exploring Cross-group Discrepancies in Calibrated Popularity for Accuracy/Fairness Trade-off Optimization. In *Proc. of MORS RecSys* (Seattle, WA, USA).

[24] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?. In *Proc. of ACM RecSys* (Amsterdam, Netherlands). 601–606.

[25] Elisabeth Lex, Dominik Kowald, and Markus Schedl. 2020. Modeling popularity and temporal drift of music genre preferences. *Transactions of the International Society for Music Information Retrieval* 3, 1 (2020).

[26] Elisabeth Lex, Dominik Kowald, Paul Seitlinger, Thi Ngoc Trang Tran, Alexander Felfernig, Markus Schedl, et al. 2021. Psychology-informed recommender systems. *Foundations and Trends in Information Retrieval* 15, 2 (2021), 134–242.

[27] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural Attentive Session-based Recommendation. In *Proc. of ACM CIKM*. ACM, 1419–1428.

[28] Taoying Li, Linlin Jin, Zebin Wu, and Yan Chen. 2019. Combined Recommendation Algorithm Based on Improved Similarity and Forgetting Curve. *MDPI Information* 10, 4 (2019).

[29] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proc. of ACM WWW* (Lyon, France). 689–698.

[30] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: Short-Term Attention/Memory Priority Model for Session-based Recommendation. In *Proc. of ACM SIGKDD*. 1831–1839.

[31] Malte Ludewig and Dietmar Jannach. 2018. Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction* 28 (2018), 331–390.

[32] Malte Ludewig, Iman Kamehkhosh, Nick Landia, and Dietmar Jannach. 2018. Effective Nearest-Neighbor Music Recommendations. In *Proc. of ACM RecSys Challenge* (Vancouver, BC, Canada). Article 3, 6 pages.

[33] Malte Ludewig, Noemi Mauro, Sara Latifi, and Dietmar Jannach. 2019. Performance Comparison of Neural and Non-Neural Approaches to Session-Based Recommendation. In *Proc. of ACM RecSys* (Copenhagen, Denmark). 462–466.

[34] Malte Ludewig, Noemi Mauro, Sara Latifi, and Dietmar Jannach. 2021. Empirical analysis of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction* 31, 1, 149–181.

[35] Alessandro B. Melchiorre, Navid Rekabsaz, Christian Ganhör, and Markus Schedl. 2022. ProtoMF: Prototype-based Matrix Factorization for Effective and Explainable Recommendations. In *Proc. of ACM RecSys* (Seattle, WA, USA). 246–256.

[36] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666.

[37] Yoon-Joo Park and Alexander Tuzhilin. 2008. The Long Tail of Recommender Systems and How to Leverage It. In *Proc. of ACM RecSys* (Lausanne, Switzerland). New York, NY, USA, 11–18.

[38] Andreas Peintner, Marta Moscati, Emilia Parada-Cabaleiro, Markus Schedl, and Eva Zangerle. 2022. Unsupervised Graph Embeddings for Session-based Recommendation with Item Features. In *Proc. of MORS RecSys*.

[39] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-aware recommender systems. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–36.

[40] Markus Reiter-Haas, Emilia Parada-Cabaleiro, Markus Schedl, Elham Motamedi, Marko Tkalcic, and Elisabeth Lex. 2021. Predicting Music Relistening Behavior Using the ACT-R Framework. In *Proc. of ACM RecSys* (Amsterdam, Netherlands). 702–707.

[41] Lei Ren. 2015. A Time-Enhanced Collaborative Filtering Approach. In *Proc. of NGCIT* (Qingdao, China). 7–10.

[42] Pengjie Ren, Zhumin Chen, Jing Li, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. RepeatNet: A Repeat Aware Neural Recommendation Machine for Session-Based Recommendation. In *Proc. of AAAI* (Honolulu, Hawaii, USA). Article 590, 8 pages.

[43] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proc. of UAI* (Montreal, Quebec, Canada). 452–461.

[44] Frank E. Ritter, Farnaz Tehranchi, and Jacob D. Oury. 2018. ACT-R: A cognitive architecture for modeling cognition. *WIREs Cognitive Science* 10, 3 (2018), e1488.

[45] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *Proc. of ACM CHIIR* (Regensburg, Germany). 337–341.

[46] Markus Schedl, Emilia Gómez, and Julián Urbano. 2014. Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends in Information Retrieval* 8, 2–3 (2014), 127–261.

[47] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2022. *Music Recommendation Systems: Techniques, Use Cases, and Challenges*. Springer US, New York, NY, 927–972.

[48] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018),

95–116.

[49] Yong Kiam Tan, Xinxing Xu, and Yong Liu. 2016. Improved Recurrent Neural Networks for Session-Based Recommendations. In *Proc. of DLRS* (Boston MA USA). 17–22.

[50] Baocheng Wang and Wentao Cai. 2020. Knowledge-Enhanced Graph Neural Networks for Sequential Recommendation. *MDPI Information* 11, 8 (2020), 388.

[51] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-Based Recommendation with Graph Neural Networks. *Proc. of AAAI* 33, 01, 346–353.

[52] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S. Sheng, Jiajie Xu, Fuzhen Zhuang, Junhua Fang, and Xiaofang Zhou. 2019. Graph Contextualized Self-Attention Network for Session-based Recommendation. In *Proc. of IJCAI*. 3940–3946.

[53] Liangliang Zhao, Jiajin Huang, and Ning Zhong. 2014. A context-aware recommender system with a cognition inspired model. In *Proc. of RSKT* (Shanghai, China). 613–622.