X4SR: Post-Hoc Explanations for Session-based Recommendations

Jyoti Narwariya[1], Priyanka Gupta[1], Garima Gupta[1], Lovekesh Vig[1] and
Gautam Shroff[1]

[1] TCS Research, New Delhi, India

Abstract

Session-based recommendation (SR) approaches have extensively employed deep neural networks (DNN) to provide high-quality recommendations based on a user's current interactions and item features. However, these approaches are black-box models, providing recommendations that are not understandable to the users and system designers. To further trust and transparency in the recommendation system and extrapolate insights into customer behavior, it is essential to provide explanations for why an item is recommended to a certain user. In this paper, we propose a novel post-hoc explainability method that provides explanations for a benchmark recommendation system NISER[1] at two levels; (i) Local explanations, where the method provides explanations for why an item is recommended in the current session, and (ii) Global explanations, where the method provides explanation for why an item is recommended in general across all sessions. Our method utilizes the learned item and session embeddings from the recommendation model in order to determine the most influential items for a recommendation. In contrast to using proxy models like LIME[2] or SHAP[3], utilizing the same model embeddings that were used for recommendations ensures that the explanations generated are of high fidelity and reflect the models' true behavior. Through quantitative evaluation on two publicly available datasets, we demonstrate that our approach is able to generate quality explanations in terms of salient items for a recommendation. To the best of our knowledge, our method is the first to provide a quantitative evaluation in terms of commonly used metrics for recommendation systems. We also demonstrate the value of providing verbalized explanations for various examples using LLMs[1] to improve readability of explanations.

Keywords

Session-based Recommendation, Explainable Recommendation, Post-hoc explanations, Verbalized explanations, LLMs

1. Introduction

Recommendation systems (RS) are an integral component of e-commerce, online advertising and streaming applications allowing systems to provide relevant content, boost sales and improve user experience. Our interest lies in SR systems [4, 5, 6, 7, 1] where the system has to dynamically make recommendations based on current session interactions without any prior user history. Nearest neighbour-based approaches like STAN[4] recommend items based on similar prior sessions. Such methods are understandable and provide reasonable explanations.

[1] we used OpenAI GPT-3 (https://openai.com/api/) to obtain verbalized explanations

eCom'23: ACM SIGIR Workshop on eCommerce, July 27, 2023, Taipei, Taiwan

jyoti.narwariya@tcs.com (J. Narwariya); priyanka.g35@tcs.com (P. Gupta); garimagupta26.work@gmail.com (G. Gupta); lovekesh.vig@tcs.com (L. Vig); gautam.shroff@tcs.com (G. Shroff)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

However, there appears to be a trade-off between a model's ability to learn complex user behavior and its interpretability. Modern recommendation systems utilize high dimensional latent features, i.e. item or session embeddings to achieve state-of-the-art performance [5, 6, 7, 1]. DNN-based SR approaches provide high-quality recommendations based on the user's current interactions and the items' latent features, but at the cost of interpretability, trust, and transparency. Generating explanations along with recommendations is essential to build trust, and improve user satisfaction, while assisting system designers to rectify irrelevant recommendations [8].

In order to make DNNs interpretable, one prominent technique employed is to learn a less complex proxy model to locally mimic and understand a DNNs behavior [2, 3]. However, this requires additional training and the explanations are not guaranteed to mimic the exact pattern of reasoning in the SR model. Another post-hoc approach [9] generates personalized post-hoc explanations based on item-level causal rules to explain the behaviors of a sequential recommendation model. However, it compromises on recommendation accuracy by constraining the model to rely on the causal rules. CGSR [10] provided explanations on session and item levels by generating a set of scores, i.e., causality and correlation scores. However, it can not be applied to any other SR approach due to unavailability of causality scores. SSR [11] generates explanations within a session by considering three factors: sequential patterns, repetition clicks, and item similarities. However, it does not provide explanations at aggregate level (i.e. global explanations). In contrast to using a proxy model or rules, our approach does not require additional training nor compromises on recommendation accuracy. We propose using the session and item representations learned from a DNN-based SR model to generate explanations that are of high fidelity, are trustworthy, and reflect the models' true behavior. Toward this, we propose **X4SR**: Post-hoc E**x**planations for **S**ession-based **R**ecommendations. We demonstrate the enhanced ability of **X4SR** to generate quality explanations in terms of explaining items at two levels: **Local explanations**: explanation of recommended items for the current session, and **Global explanations**: explanations for the recommended item at an aggregate level. Local explanations are important for end-users/customers to trust the system, and global explanations are useful for a business user to understand aggregated customer behavior. The generated explaining items along with meta-information, i.e. current session (in case of local explanations) and similar prior sessions (in case of local and global explanation) can be parsed and reasoned over via LLMs to get verbalized explanations that are understandable to the user as well as the system designer. While several explanation approaches have been proposed in the literature, none of them evaluate the generated explanations quantitatively. **X4RS** provides quantitative evaluation in terms of commonly used metrics in RS such as Recall and MRR.

In this work, we employ NISER [1], a well known session based recommendation benchmark to validate our explainability approach, though it should be noted that **X4RS** is model agnostic and can be employed for any embedding based SR model. We summarized the key contributions as follows: (i) We propose a post-hoc method to generate explanations that reflect the models' true behavior at two levels: Local and Global, (ii) To the best of our knowledge, our approach is first to provide a quantitative evaluation in terms of commonly used metrics such as Recall and MRR, and (iii) We provide verbalized explanations via LLMs to improve the readability of explanations.

2. Proposed Approach

2.1. Problem Setting

Let $\mathcal{S}_{tr}$ and $\mathcal{S}_{te}$ be the set of prior (train) sessions and current (test) sessions, respectively. We consider $\mathcal{I}$ to be the set of m items observed in set $\mathcal{S}_{tr}$. Given any current session $s \in \mathcal{S}_{te}$, which is a sequence of l item-click events, $\mathcal{I}_s$ = $\{i_{s,1}, i_{s,2}, \ldots, i_{s,l}\}$, where $i_{s,j} \in \mathcal{I}$, the SR model $\mathcal{M}$ predicts a recommendation list of top k items, $\mathcal{I}_r = \{i_1, i_2, \ldots, i_k\} \subset \mathcal{I}$. From a trained SR model $\mathcal{M}$, we obtain learned item embedding $\mathbf{i}_j \in \mathbb{R}^d$ for each item of $\mathcal{I}$ and denote the item embedding set as **I**. Similarly, we obtain learned session embedding as $\mathbf{s} \in \mathbb{R}^d$ for all prior and current sessions denoted by $\mathbf{S}_{tr}$ and $\mathbf{S}_{te}$, respectively. In this work, we consider the benchmark SR model NISER [1] as $\mathcal{M}$.

The goal for explainable SR approaches is to explain each recommended item $i_j \in \mathcal{I}_r$ at the session level (why item i_j is recommended in current session $s \in \mathcal{S}_{te}$), as well as at a global level (why item i_j is recommended across all user sessions).

2.2. X4RS: Post-hoc Explanation

X4RS generates explanations in terms of explaining items for each recommendation using learned latent embeddings of sessions and items $\mathbf{S}_{tr}$, $\mathbf{S}_{te}$ and **I**. The explanations are given at two levels:

A) **Local Explanations:** To generate explanations for a recommended item $i_j \in \mathcal{I}_r$ for session s, we first compute cosine similarity between the current session embeddings **s** and prior session embeddings for sessions where i_j is present in their history $\mathbf{S}^j_{tr}$. We thus obtain the top-n most similar candidate prior sessions $\mathcal{S}^{s,j}_{tr}$. All items in prior candidate sessions are added to the candidate items set $\mathcal{I}^{s,j}_{tr}$. Then, we obtain relevant items $\mathcal{X}^{s,j}_{tr}$ from the candidate items set based on: i) pair-wise similarity between all the items in candidate item set, items-pair with similarity greater than equal to threshold β, ii) items that occur most frequently in the candidate items set. Further, explaining items from session s are the ones having maximum similarity with the relevant items. We summarize the process in Algorithm 1.

B) **Global Explanations:** To generate generalized explanations for item i_j, we first obtain all the prior sessions in which i_j is present in the session history as $\mathcal{S}^j_{tr}$, and cluster them using DBSCAN [12]. We consider density-based clustering i.e. DBSCAN instead of distance-based clustering e.g., K-Means because it allows to learn clusters of arbitrary shape with no prior knowledge of number of clusters. We estimate centroid of each cluster as average of embeddings of sessions present in respective cluster. For each cluster c, we obtain top-n candidate prior sessions $\mathcal{S}^{c,j}_{tr}$ that are most similar to the centroid of the cluster c. Further, we consider set of all items that are present in candidate prior sessions as candidate items set $\mathcal{I}^{c,j}_{tr}$. Next, we obtain explaining items based on pair-wise similarity between all the items in the candidate items set (similarity $\geq \beta$) and most frequent items in candidate item set. The explaining items for each cluster corresponds to different user behavior patterns, which shows that different metapath are responsible for positive interaction (e.g. click, buy, etc.) on item i_j. We summarize the process in Algorithm 2.

Algorithm 1 Local explanations

```
Given recommended items $\mathcal{I}_r$, Item clicked history in session s as $\mathcal{I}_s$, learned item embeddings $\mathbf{I}$, prior and current sessions embeddings $\mathbf{S}_{tr}$ and $\mathbf{S}_{te}$.
for each current session $s$ in $\mathcal{S}_{te}$ do
    for each recommended item, $i_j \in \mathcal{I}_r = \{i_1, i_2, ..., i_k\}$ do
        $\mathcal{S}^{j}_{tr} = \{s' \mid i_j \in \mathcal{I}_{s'}\}, \forall s' \in \mathcal{S}_{tr}$
        Candidate prior sessions $\mathcal{S}^{s,j}_{tr} = \arg\max_n(cosine(\mathbf{S}^{j}_{tr}, \mathbf{s}))$
        Candidate items set $\mathcal{I}^{s,j}_{tr} = \cup \{i' \in \mathcal{I}_{s'}\}, \forall s' \in \mathcal{S}^{s,j}_{tr}$
        $I_1$ = Most frequent items across $\mathcal{S}^{s,j}_{tr}$, $I_1 \subset \mathcal{I}^{s,j}_{tr}$
        $I_2$ = {i' | $max(cosine(\mathbf{i'}, \mathbf{I}^{s,j}_{tr})) \geq \beta$ }, $\forall i' \in \mathcal{I}^{s,j}_{tr}$
        Relevant items $\mathcal{X}^{s,j}_{tr} = I_1 \cup I_2 - i_j$
        $Sim^{s,j} = cosine(\mathbf{I}_s, (\mathbf{X}^{s,j}_{tr} \oplus \mathbf{i}_j))$        ▷ $\oplus$ : concatenation
        $x^{s,j}_{te} = \arg\max_{i' \in \mathcal{I}_s} Sim^{s,j}$
    end for
    Explaining items for session s, $\mathcal{X}^{s}_{te}$ = Most frequent items from $x^{s,j}_{te}$ $\forall$ $i_j \in \mathcal{I}_r$ ▷ Here, we consider top-2 frequency for selecting explaining items in session s.
end for
```

Algorithm 2 Global explanations

```
Given recommended items $\mathcal{I}_r$, Item clicked history in session s as $\mathcal{I}_s$, learned item embeddings $\mathbf{I}$, prior and current sessions embeddings $\mathbf{S}_{tr}$ and $\mathbf{S}_{te}$.
for Each item $j = 1, 2, \ldots, m \in I$ do
    $\mathcal{S}^{j}_{tr} = \{s' \mid i_j \in \mathcal{I}_{s'}\}, \forall s' \in \mathcal{S}_{tr}$
    Clusters $\mathcal{C}$ = DBSCAN($\mathbf{S}^{j}_{tr}, \epsilon, min\_samples$)
    for c in clusters $\mathcal{C}$ do
        Centroid $\mathbf{c} = mean(\mathbf{S}^{c,j}_{tr})$, where $\mathcal{S}^{c,j}_{tr} \in c$
        Candidate prior session $S^{c,j}_{tr}$ = $\arg\max_n(cosine(\mathbf{S}^{j}_{tr}, \mathbf{c}))$
        Candidate items set $\mathcal{I}^{c,j}_{tr} = \cup \{i' \in \mathcal{I}_{s'}\}, \forall s' \in \mathcal{S}^{c,j}_{tr}$
        $I_1$ = Most frequent item across $S^{c,j}_{tr}$, $I_1 \subset S^{c,j}_{tr}$
        $I_2$ = {i' | $max(cosine(\mathbf{i'}, \mathbf{I}^{c,j}_{tr})) \geq \beta$ }, $\forall i' \in \mathcal{I}^{c,j}_{tr}$
        Explaining items $X^{c,j}_{tr} = I_1 \cup I_2$ - $i_j$
    end for
end for
```

3. Experimental Evaluation

3.1. Dataset Details

We evaluate efficacy of **X4RS** on two publicly available datasets, i.e. Diginetica (DN) and Amazon Musical Instruments (AMI). The DN[1] dataset is a large-scale real-world transactional data from CIKM Cup 2016 challenge. The AMI ratings dataset[2] is a public dataset from Amazon, which contains timestamped user-item interactions from May 1996 to Oct 2014 and metadata contains items' descriptions, categories, brands, etc. We follow [7] for dataset pre-processing. For **DN**, we filter out items which have frequency less than 5, followed by removal of sessions

[1] https://competitions.codalab.org/competitions/11161
[2] https://nijianmo.github.io/amazon/index.html

of length 1. We consider sessions from last 1 week as test data. Finally, we consider $0.7M|$ $30,574$ sessions for training| testing with average session length 5.12 and $43,097$ items. For **AMI**, we consider most frequent $10k$ users data, remove items with frequency less than 5. We consider user's transactions (i.e. users' ratings) lying within 20 minutes as a session. In addition, the sessions from last 1 day are used as the test data for AMI. Finally, we consider $18,128|$ $6,126$ sessions for training| testing with average session length 6.45 and $2,451$ items. For both datasets, we split the remaining data chronologically as a training set and validation set for training and model selection purposes respectively. We filtered out all sessions of length less than 3 from testing data for explanation that allows to obtain atleast 2 explaining items in the session.

3.2. Evaluation

We consider two evaluation settings: quantitative, and qualitative. For quantitative evaluation, we remove/replace the explaining items generated by **X4RS** from test sessions and observe the performance of the SR model. The idea is to validate if explaining items are necessary for the SR model to recommend the item that was actually clicked or bought in the original test session. We compare performance on: i) original test sessions (OTS), ii) by removing non-explaining items (-NX), iii) by removing items based on popularity index (-P), where popularity index of an item is calculated by dividing total sales/clicks of the item by total sales/clicks of all items, iv) by removing the explaining items (-X) obtained from our approach, v) by replacing the specific items with items that are at the highest distance based on cosine similarity, i.e., replacing explaining items (-X+F), and replacing popular items (-P+F). We use the standard evaluation metrics Recall (R@K) and Mean Reciprocal Rank (MRR@K) as used in [7, 1]. R@K represents the proportion of test instances which has target item in the top-K items. MRR@K is the average of reciprocal ranks of target item in the recommendation list. For qualitative evaluation, we input explaining items along with metadata, current test sessions, and candidate prior sessions to GPT-3 and obtain verbalized explanations that are understandable by business users and system designers. We conducted a user study to validate the responses obtained from GPT-3. We received feedback scores between 1 and 5 (higher the score better the quality of generated text) from 20 users for 10 explanations each for both the levels, i.e., local and global. **Note:** Post-hoc approaches [2, 3, 9, 11] in literature utilize items' features for providing explanations that are not available in this work. Therefore, we are unable to compare the proposed approach with these approaches and consider alternate procedures to baseline.

Hyperparameter Setup We use a hold-out validation set for model selection using Recall (R@20) as the performance metric for all experiments in Table 2. Following [1], we use $d = 100$ and a learning rate of 0.001 with the Adam optimizer. We employ grid-search over β in $\{0.65, 0.5, 0.4, 0.3, 0.25\}$. The best parameters on the validation set are $\beta = 0.5$ and 0.25 for DN and AMI, respectively. While explaining items at the global level, we used $\epsilon = 0.001$ and $min\_samples = 4$ for $|\mathcal{S}_{tr}^{j}| >= 20$ otherwise $min\_samples = 2$ are the best on the validation set. We use $n = 5$ for obtaining candidate prior sessions.

Table 1
Global explainability evaluation. Test sessions are considered where respective item is recommended. We selected one item each from long-tail (less popular), mid (moderate popular), head (more popular) randomly for both the datasets. Here, OTS: Original test sessions, X: Explaining items, NX: Non-explaining items, F: Farthest items, P: Popular items. Best results, i.e. percentage drop (% ↓) in Recall (R@20) and MRR (MRR@20) are marked in bold and second set of results are marked as bold-italic.

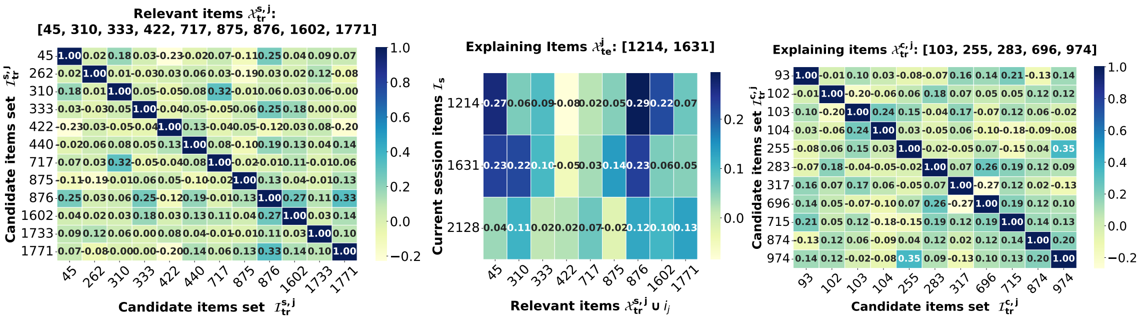
Test-	DN						AMI					
sessions	Long-tail(Item: 17009)		Mid (Item: 2125)		Head (Item: 94)		Long-tail (Item: 175)		Mid (Item: 39)		Head (Item: 102)	
variants	R@20 (% ↓)	MRR@20 (% ↓)	R@20 (% ↓)	MRR@20 (% ↓)	R@20 (% ↓)	MRR@20 (% ↓)	R@20 (% ↓)	MRR@20 (% ↓)	R@20 (% ↓)	MRR@20 (% ↓)	R@20 (% ↓)	MRR@20 (% ↓)
OTS	12.50	1.79	41.82	14.45	68.42	20.41	12.00	7.00	30.32	19.41	42.94	30.62
-NX	12.50 (0%)	0.96 (-46%)	39.09 (-7%)	12.20 (-16%)	57.89 (-15%)	38.16 (87%)	12.00 (0%)	4.57 (-35%)	28.53 (-6%)	16.67 (-14%)	40.15 (-6%)	25.50 (-17%)
-P	**0.00 (-100%)**	**0.00 (-100%)**	31.82 (-24%)	11.30 (-22%)	**42.11 (-38%)**	**7.98 (-61%)**	12.00 (0%)	3.57 (-49%)	**23.43 (-23%)**	**10.45 (-46%)**	**37.07 (-14%)**	**20.90 (-32%)**
-X	**0.00 (-100%)**	**0.00 (-100%)**	**22.73 (-46%)**	**5.90 (-59%)**	47.37 (-31%)	15.45 (-24%)	**8.00 (-33%)**	**0.60 (-91%)**	24.33 (-20%)	13.63 (-30%)	37.72 (-12%)	22.94 (-25%)
-P+F	**0.00 (-100%)**	**0.00 (-100%)**	**11.82 (-72%)**	**4.43 (-69%)**	**21.05 (-69%)**	**7.07 (-65%)**	***0.00 (-100%)***	***0.00 (-100%)***	***16.10 (-47%)***	***5.63 (-70%)***	29.83 (-30%)	***12.01 (-60%)***
-X+F	**0.00 (-100%)**	**0.00 (-100%)**	**11.82 (-72%)**	**4.43 (-69%)**	**21.05 (-69%)**	**7.07 (-65%)**	8.00 (-33%)	1.02 (-85%)	16.99 (-44%)	8.78 (-55%)	**27.32 (-36%)**	12.63 (-59%)

Table 2
Local explainability evaluation. Here, OTS: Original test sessions, X: Explaining items, NX: Non-explaining items, F: Farthest items, P: Popular items. Best results, i.e. drop in i.e. percentage drop (% ↓) in Recall (R@20) and MRR (MRR@20) are marked in bold and second set of results are marked as bold-italic.

Test-sessions	DN		AMI	
variants	R@20 (% ↓)	MRR@20 (% ↓)	R@20 (% ↓)	MRR@20 (% ↓)
OTS	44.16	12.55	26.64	16.81
-NX	43.70 (-1%)	12.42 (-1%)	25.89 (-3%)	16.31 (-3%)
-P	40.26 (-9%)	11.20 (-11%)	20.81 (-22%)	10.91 (-35%)
-X	**36.90 (-16%)**	**10.38 (-17%)**	**18.89 (-29%)**	**9.87 (-41%)**
-P+F	22.77(-48%)	6.51(-48%)	15.80 (-41%)	6.41(-62%)
-X+F	***15.92 (-64%)***	***4.40 (-65%)***	***8.68 (-67%)***	***3.07 (-82%)***

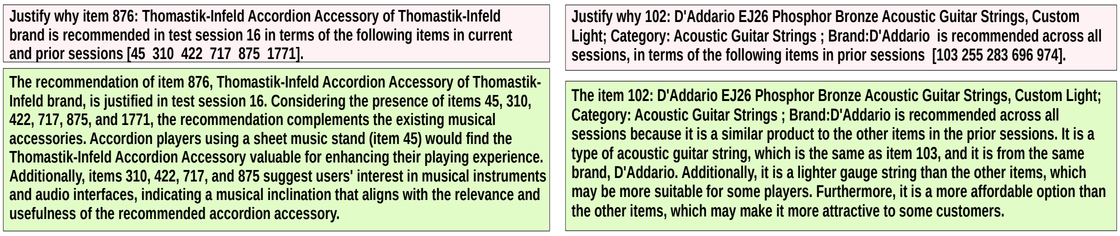
3.3. Results and Observations

Quantitative: Table 2 and 1 show the performance for local and global explanation, respectively. From table 2, we observe that by removing non-explaining items from test sessions (-NX), Recall@20 (R@20) and MRR@20 are dropped by 1% and 3% as compared to OTS for DN and AMI, respectively. Slight percentage drops indicate that non-explaining items are irrelevant to recommend the target item. Further, if popular items are removed (-P), we observe considerable drops i.e., 9% and 22% in R@20, and 11% and 35% in MRR@20, indicating popular items are relevant. However, we observe significant percentage drops when explaining items are removed (-X) i.e, R@20 by 16% and 29% and MRR@20 by 17% and 41% for DN and AMI, respectively. This indicates that explaining items generated by our approach are crucial to recommend the target item. Moreover, we observe a further drop in R@20 by 64% and 67%, and MRR@20 by 65% and 82% if explaining items are replaced with the least similar items out of all the items (-X+F) due to additional noise. Also, drop in R@20 by 48% and 41%, MRR@20 by 48% and 62% if popular items are replaced instead of explaining items. However, drops are better in case of replacing explaining items. This further validates the efficiency of our approach to generate explaining items. Similarly, from table 1, we observe significant percentage drops while removing explaining items (-X) in terms of R@20 as 100%, 46%, 31% and 33%, 20%, 12% for Long-tail, Mid, Head item for DN and AMI, respectively that is significantly better than removing non-explaining items (-NX) i.e., 0%, 7%, 15% and 0%, 6%, 7%. Also, it is comparable



(a) (b) (c)

Figure 1: (a,b) provide local explanations: why *Item 876* is recommended in session *16*, (a) provides relevant items based on pair-wise similarity and frequency, (b) obtains explaining items using relevant items and (c) global explanations for *Item 102*.



Justify why item 876: Thomastik-Infeld Accordion Accessory of Thomastik-Infeld brand is recommended in test session 16 in terms of the following items in current and prior sessions [45 310 422 717 875 1771].

The recommendation of item 876, Thomastik-Infeld Accordion Accessory of Thomastik-Infeld brand, is justified in test session 16. Considering the presence of items 45, 310, 422, 717, 875, and 1771, the recommendation complements the existing musical accessories. Accordion players using a sheet music stand (item 45) would find the Thomastik-Infeld Accordion Accessory valuable for enhancing their playing experience. Additionally, items 310, 422, 717, and 875 suggest users' interest in musical instruments and audio interfaces, indicating a musical inclination that aligns with the relevance and usefulness of the recommended accordion accessory.

Justify why 102: D'Addario EJ26 Phosphor Bronze Acoustic Guitar Strings, Custom Light; Category: Acoustic Guitar Strings ; Brand:D'Addario is recommended across all sessions, in terms of the following items in prior sessions [103 255 283 696 974].

The item 102: D'Addario EJ26 Phosphor Bronze Acoustic Guitar Strings, Custom Light; Category: Acoustic Guitar Strings ; Brand:D'Addario is recommended across all sessions because it is a similar product to the other items in the prior sessions. It is a type of acoustic guitar string, which is the same as item 103, and it is from the same brand, D'Addario. Additionally, it is a lighter gauge string than the other items, which may be more suitable for some players. Furthermore, it is a more affordable option than the other items, which may make it more attractive to some customers.

(a) Explanations for recommended item *876* in session 16

(b) Explanations for recommended item *102* across sessions

Figure 2: Verbalized explanations using openAI GPT-3

by removing popular items (-P) i.e., 100%, 24%, 38% and 0%, 23%, 14%. Moreover, when explaining items are replaced with the least similar items out of all the items (-X+F), the drops in R@20 is as significant as 100%, 72%, 69%, and 33%, 44%, 36% for DN and AMI, respectively. We also observed similar drops if popular items are replaced instead, i.e., drops in R@20 are 100%, 72%, 69%, and 100%, 47%, 30% for DN and AMI, respectively. Similar percentage drops are observed for MRR@20.

Qualitative Analysis: Case Study on Amazon Musical Instrument Dataset: We consider Amazon dataset due to the availability of the meta information of the items, which is not the case with Diginetica. First, we study why item "***876***: Thomastik-Infeld Accordion Accessory" of brand "Thomastik-Infeld" is recommended in a session 16. Figure 1a shows pair-wise similarity between candidate items set. The pairs with high similarity, i.e. 0.33 and 0.32 are [876, 1771] and [717, 310], respectively. Hence, relevant items based on similarity are '***1771***: Line 6 Relay G50 Wireless Guitar System', '***717***: Pedaltrain MINI With Soft Case, Instrument Cable; Stage & Studio Cables; and '***310***: Fender F Neckplate Chrome'. The relevant items based on frequency are as follows: "***45***: On-Stage Professional Grade Folding Orchestral Sheet Music Stand', '***310***: Fender F Neckplate Chrome', '***422***: Classic Series Instrument Cable with Right Angle Plug', '***875***: Behringer Ultimate Guitar-to-USB Audio Interface'. Further, similarity between relevant items and current session items is shown in figure 1b. We observe that explaining items in current session '***1214***: Fender Precision Bass Pickups' and '***1631***: Electric Guitar Bass Pickguard Screws' is close to item ***45***, item ***310*** and recommend item ***876***. This is because they belong to the guitar accessories category. *From figure 1a and 1b, we can conclude that item **876** is recommended in session* 16 *because explaining items and relevant items are related to musical accessories. We obtained the similar verbalized explanations from GPT-3 as shown in figure 2a.*

Further, we study why an item "***102***: Phosphor Bronze Acoustic Guitar Strings, Custom Light, which belongs to the 'Acoustic Guitar Strings' category and 'D'Addario' brand is recommended in general. Figure 1c shows that it is similar to '***255***: Planet Waves Acoustic Guitar Quick-Release System', '***974***: Martin M Acoustic Guitar Bridge Pins', '***283***: Snark SN1 Guitar Tuner', '***696***: Ernie Ball Earthwood Light Phosphor Bronze Acoustic String Set' and '***103***: D'Addario Phosphor Bronze Acoustic Guitar Strings, Medium' with similarities 0.35, 0.35, 0.26 and 0.24 respectively, i.e. all explaining items related to guitar accessories and all relevant prior sessions contains same brand 'D'Addario' items. We observe similar explanations from GPT-3 as shown in figure 2b that *item* 102 *is in the same category (Acoustic Guitar Strings) and same brand (D'Addario) as the explaining item* 103. *Additionally, it is a lighter gauge string than the other items, which may be more suitable for some customers.*

Further, we evaluate verbalised explanations using feedback scores from user-study. 67% of users gave 4 and 5 scores for all global and local explanations and around 20% of users gave a score of 3 for these verbalised explanations. On average local and global explainability scores (out of 5) are 3.85 and 4.13, respectively.

4. Discussion

We highlighted the issue of trust and transparency in benchmark DNN-based recommendation systems such as NISER[1]. We showed that in contrast to learn proxy models like [2, 3], which require additional training, **X4RS** use learned item and session embeddings from NISER to generate high fidelity, trustworthy explanations. We observed a significant drop in Recall (R@20) and MRR@20 after removing explaining items from test sessions that validate the quality of the explanations. Further, we observed that verbalized explanations obtained from GPT-3 improved the readability of explanations for users and system designers. In future, we would like to explore **X4RS** with other approaches such as NARM[13], GRU4Rec[5], STAMP[6], SASRec [14] and CL4Rec[15].

References

[1] P. Gupta, D. Garg, P. Malhotra, L. Vig, G. Shroff, Niser: Normalized item and session representations with graph neural networks, arXiv preprint arXiv:1909.04276 (2019).

[2] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[3] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[4] D. Garg, P. Gupta, P. Malhotra, L. Vig, G. Shroff, Sequence and time aware neighborhood for session-based recommendations: Stan, in: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 2019, pp. 1069–1072.

[5] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, arXiv preprint arXiv:1511.06939 (2015).

[6] Q. Liu, Y. Zeng, R. Mokhosi, H. Zhang, Stamp: short-term attention/memory priority model for session-based recommendation, in: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 1831–1839.

[7] S. Wu, Y. Tang, Y. Zhu, L. Wang, X. Xie, T. Tan, Session-based recommendation with graph neural networks, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019.

[8] Y. Zhang, X. Chen, et al., Explainable recommendation: A survey and new perspectives, Foundations and Trends® in Information Retrieval 14 (2020) 1–101.

[9] S. Xu, Y. Li, S. Liu, Z. Fu, X. Chen, Y. Zhang, Learning post-hoc causal explanations for recommendation, arXiv preprint arXiv:2006.16977 (2020).

[10] C. Geng, H. Wu, H. Fang, Causality and correlation graph modeling for effective and explainable session-based recommendation, arXiv preprint arXiv:2201.10782 (2022).

[11] J. Chen, W. Wu, W. Hu, W. Zheng, L. He, Ssr: Explainable session-based recommendation, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.

[12] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: kdd, volume 96, 1996, pp. 226–231.

[13] J. Li, P. Ren, Z. Chen, Z. Ren, T. Lian, J. Ma, Neural attentive session-based recommendation, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1419–1428.

[14] W.-C. Kang, J. McAuley, Self-attentive sequential recommendation, in: 2018 IEEE international conference on data mining (ICDM), IEEE, 2018, pp. 197–206.

[15] X. Xie, F. Sun, Z. Liu, S. Wu, J. Gao, J. Zhang, B. Ding, B. Cui, Contrastive learning for sequential recommendation, in: 2022 IEEE 38th international conference on data engineering (ICDE), IEEE, 2022, pp. 1259–1273.