Computer Science
&
Data Science

Capstone Report - Spring 2022

User Intention Detection and Evaluation for Recommendation Systems

Yuan Huang,
Ziyi Jiang,
Mingqian Zheng

Supervised by
Hongyi Wen

Preface

This project is the final output of the NYUSH Computer Science and Data Science capstone project to partially fulfill the degree requirement. Recommendation system is a promising and recently developed area in computer science, where most of the studies focus on inventing recommendation algorithms but how to better evaluate the system performance remains a tricky question to be answered. We noticed that intention-aware metrics lie in the vacancy of the study area. Hence, in this project, we used the MovieLens 1M dataset to first extract users intention and then evaluate the performance of commonly used recommendation models on our self-defined intention-based recall. The experiment results point out that our metric is able to provide a new angle to evaluate recommendation models compared with other traditional metrics that are widely-used for information retrieval systems in general.

Acknowledgements

We would like to thank our supervisor, Prof. Wen Hongyi, for his professional guidance and unlimited support throughout the process. We are also grateful to the capstone instructors Prof. Guo Li and Prof. Olivier Marin. Finally, we would like to express our gratitude to our families and friends for their continuous support.

Abstract

Within the field of recommendation systems, most previous studies aimed to improve the accuracy of recommendation models. Recently, user intention has been a promising research direction for better model performance. However, the commonly used metrics focus exclusively on the measurement of accuracy. In this study, we proposed an intention-aware recall metric along with a statistical mechanics of user intention extraction. By comparing four distinct models, we have verified that our metric does capture information that is usually ignored by traditional metrics, thus providing a new aspect to evaluate recommendation models.

Keywords

User Intention Detection; Recommendation System; Metric Design

Contents

1 Introduction

The explosively high volume of today's digital information leaves us with the problem of information overload. To address this challenge, recommendation systems are developed to retrieve the most relevant information for users, allowing them to get timely access to satisfy their need for personal information. A crucial part to consider when developing recommendation systems is understanding users' preferences and the underlying intention. In contrast to inherent user preference, which is often static and constant in traditional recommendation algorithms, user intention varies dynamically in different settings [1]. Extensive research has studied how to predict user-item interactions with high accuracy [2, 3]. In contrast, user intention is a relatively new study area, and only a few recent works have started to work on building models to extract users' interests and intentions [4, 5]. Up till now, there is no consensus among researchers on how to measure recommendation performance with respect to user intention, and the existing studies on user intention still depend on the traditional evaluation metrics in information retrieval systems, including recall, hit rate, NDCG, etc.

In this project, we propose an intention-based recall metric to evaluate the alignment of the recommended list with user intention. We conducted statistical analysis to extract user intention as a preliminary step since we lack ground-truth user intention due to limitations in data collection. We compared four distinct models to evaluate the metric to verify its reliability and interpretability. Our metric can be applied to both short-term and long-term user intention with no additional requirements for the dataset. We demonstrate that this new intention-based recall metric is able to reflect the differences in performance of various models in capturing user intention. Finally, we discuss the limitations of analyzing user intention in offline evaluations and suggest a few directions for future research.

2 Related Work

2.1 Modeling User Intention

To predict interaction between users and items, traditional recommendation techniques mainly utilize Collaborative Filtering (CF) that depends solely on prior user behavior without requiring explicit user profiles [6, 7]. By analyzing the similarity between users/items and user-item interactions, CF is able to predict likely user behavior [6]. Typically, it can be divided into two types: Memory-Based and Model-Based. The Memory-Based models predict new ratings through the sum of a weighted average rating between similar users and items [8]. Model-Based techniques include Clustering techniques, Association techniques, Bayesian networks, Neural Networks, etc [9]. They utilize the pre-computed models trained through machine learning or data mining techniques and give predictions and recommendations in a short period of time. Despite this significant advantage, CF suffers from the *cold start* problem when a new item pops up with no prior information, which is another ongoing research field [2].

More recently, the recommendation task has been formalized as a sequential recommendation problem, which aims to model the context of user activities with a focus on the relative order of user behaviors as sequences [10]. To analyze the context of user behavior, it's necessary to distinguish transient effects from long-term effects in terms of user preferences [11]. In addition, user intention is another important factor that influences the observed feedback. To model user intention, Hidden Markov Model (HMM) was utilized for recognizing intentions underlying user behavior [12, 13]. Yao et al.[14] proposed a Recurrent Neural Network (RNN) method to form a language model for attention and intention processes in human conversations.

2.2 Recommendation Evaluation

The evaluation of recommendation systems has been a big challenge, considering the various goals of evaluation and the inconsistent performance of the same algorithms in different settings or domains [15]. In the early stages of development, recommendation evaluation prioritizes accuracy or relevance, which means the higher similarity between the recommended items and user consumed items indicates better performance. The most popular metrics used in information retrieval include Recall, Precision, Hit Rate, Normalized Discounted Cumulative Gain (NDCG), etc. [16, 17]. However, evaluation beyond accuracy has been of great interest since the beginning of the 2000s [15]. With state-of-the-art techniques in deep learning, such as RNN, many recommendation models can achieve a pretty high score for the sake of accuracy, which inspires further research on the "quality of recommendation" [18].

Ge et al. [18] proposed a metric that focuses on the trade-off between *coverage* and *serendipity*, which refers to the ratio of the recommended domain over the whole domain of the system and the unexpectedness of the recommended set respectively. Previous research has revealed that how to define a metric for recommendation evaluation can be very subjective, and even the same term can have several different interpretations. Amigó et al. [19] conducted a constraint-based axiomatic framework, Rank-Biased Utility (RBU), to evaluate and compare the suitability of existing metrics. As a prominent feature of recommendation, diversity is an essential criterion at the opposite of accuracy [20]. Castells et al. [21] collected distinct metrics for novelty and diversity in recommendations to give a unified view. Moreover, the results of offline and online evaluations are likely to contradict each other and should follow different criteria accordingly [22].

Although more and more studies are heading to build an intention-aware model [4], there's still no commonly applied metric for evaluations of user intention. Specifically, previous research still used metrics such as diversity and serendipity to examine model performance beyond accuracy. These metrics do not measure the alignment between recommendations and user intention. We aim to propose an intention-based Recall metric so that the metric can evaluate whether the model captures users' short-term and long-term intentions.

3 Solution

3.1 Dataset

The available datasets in recommendation systems differ in their data attributes. The data attributes can be divided into three types in general: user-specific attributes, item-specific attributes, and user-item interactions. User-specific attributes include demographic information, social network, cross-platform information, etc. Item-specific attributes include the item type, popularity, seasonal dynamics, etc. Compared with user-specific attributes, item attributes can vary a lot for different recommendation systems. As for user-item interactions, they reflect user clicks, watches, purchases, ratings, etc. To ensure the generalizability of our metric, we only use the most basic data attributes that are widely shared by all the datasets in the field of recommendation system: userID, itemID, item category, and a binary indicator of user-item interaction. In the following subsections, we first describe our chosen dataset and then introduce the method of intention extraction.

3.1.1 Overview of the dataset: MovieLens (1M)

The MovieLens datasets are commonly used to examine the recommendation system that was collected by the GroupLens Research Project at the University of Minnesota [23]. There are several versions ranging in size from 100K to 1B with changes in data attributes. In this project, we use the MovieLens 1M dataset to train and test our recommendation systems by considering the trade-off between the sample size, the information richness, and the running time. MovieLens

1M is a stable benchmark dataset released in 2003 containing more than 1 million ratings from 6039 users with 3952 movies. According to Table 1, there are 18 genres of the movies and a total of 1040 days with user ratings from 2000-04-25 to 2003-02-28.

Dataset	users	movies	rates	movie genres	rating days
MovieLens (1M)	6039	3952	1000209	18	1040

Table 1: `Dataset Overview`

For **users**, we only use the attribute user ID.

For **movies**, we have attributes for movie ID and the genres. There are 18 different genres of movies, and one movie can have multiple genres. The genres contain Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and Western.

For **user-item ratings**, we have attributes for user ID, movie ID, rating, and timestamp. The rating is treated as a binary indicator, and the timestamp is represented in seconds. One thing to be noticed is that each user has at least 20 ratings.

3.1.2 Data Discovery

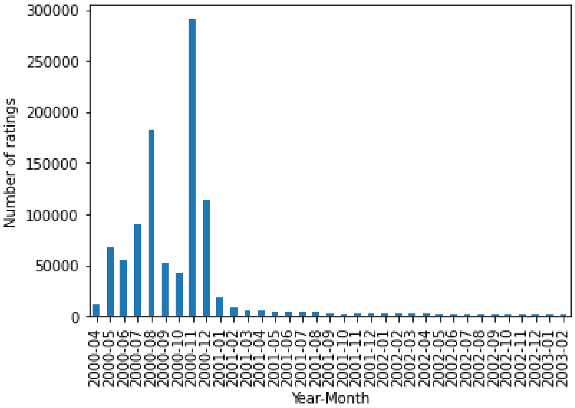

Figure 1: # user ratings in different months

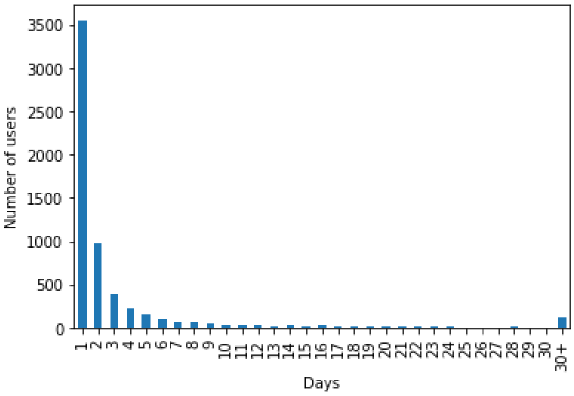

Figure 2: # users who rates in different day ranges

Rating distribution. The following figures visualize the data discovery based on the rating dataset and movie dataset.Figure 1 is the bar chart depicting the amounts of ratings given by all users each month. The right-skewed graph implies an uneven distribution among months. Specifically, ratings were quite frequent from 2000-04 to 2001-01, with few ratings following that. Figure 2 shows the time distribution of users' ratings, namely the distribution of the number of users who rated in different day ranges. We can see that more than 50 per cent of users gave all of their ratings within 1 day, and most of the remaining users completed their ratings within 2-9 days, with fewer ratings recorded over 10 days. Combining Figure 1 and Figure 2, we could see that MovieLens 1M is relatively sparse, which is consistent with what has been stated in previous studies.

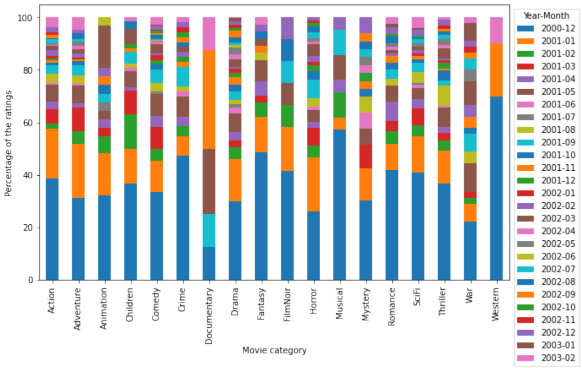

Figure 3: Percentage of the # ratings on movies category for user 195 over months

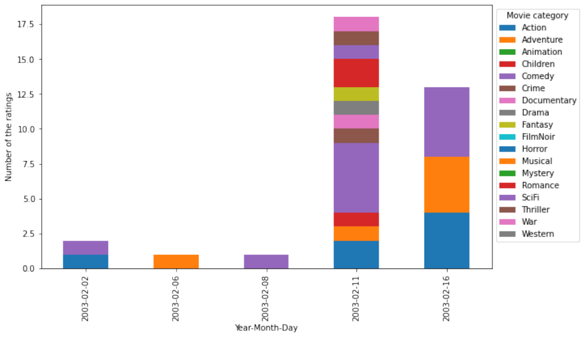

Figure 4: # ratings on movies category for user 195 within the latest month (2003-02 in this case)

Intention Analysis: A Case Study. Given the sparsity of the dataset and the uneven distribution of ratings, the intention analysis may not make sense for users who made their ratings within a short period of time, especially in extreme cases within one day. This limitation of the dataset will be further discussed later. In order to generate reasonable user intention, we took an "ideal" user as an example to develop the general algorithm. The ideal case is that a user's ratings are distributed over months so that it makes sense to talk about the dynamic of his/her long-term intention. For these reasons, we targeted user 195 as a representative example for a case study considering his rating spans the longest period of time.

Figure 3 and Figure 4 give us an overview of how user 195's movie interest changed throughout two years. Figure 3 displays the time distribution of ratings for each movie category over the span of 27 months. When we compare vertically, we can see that the user has a lot of ratings in the time period starting from 2000, whereas there are few ratings in 2002-2003. As for horizontal comparison, each bar represents the time distribution of ratings for one movie category. That is, the more slices of the bar chart, the more consistent user's intention toward that category. If a user rated one movie category almost every month, then we can infer that such category is a long-term user intention.

Figure 4 shows the user rating records for different categories of movies in the latest month, i.e., 2003-02, which can be used to infer his/her short-term intentions. For example, within the 5 days of rating, user 195 rated comedy on 4 days, action and adventure on 3 days. Correspondingly, we can also see in Figure 3 that user 195 has a long-term preference for comedy, action, and adventure.

3.2 User Intention Extraction

Generally speaking, we extract two types of intention from users, one is *long-term* intention, and the other is *short-term* intention. Firstly, we sort the rating records of different users by timestamps, which is accurate to the second in this case. Secondly, we divide each user's rating history into sessions of various lengths. We denote the session by s_r where r is the number of data points belonging to one session. For instance, s_5 refers to one session with 5 ratings for each user. Thirdly, in order to exclude the short-term interest from long-term interest, we take the session of long-term interest from the first to the second latest session, thus excluding the most recent session. The corresponding short-term interest is generated from the last session. Finally, we count the ratings of the same user for each movie category in the specified range of sessions, respectively, and take the first m as the short-term and long-term intentions of that user. In our study, m is set to be 3, which can be changed up to different scenarios. This way, we get the first m intentions of different users for the short-term and long-term, and the generated intentions will be applied to the evaluation process, which will be described in detail in section 4.1.

3.3 Baseline Models

We utilize three baseline models and one innovative model invented by [4] to test our proposed metric. This section will present the key idea and advantages of each model, along with our hypothesis regarding their performance in our intention-based metrics.

3.3.1 Deep Neural Network (DNN) for YouTube

YouTube DNN is a two-stage deep collaborative filtering model consisting of two deep neural networks: one for *candidate generation* and the other for *ranking* [24]. In the stage of candidate generation, we input a set of user behavior into the network so that the DNN can learn user embeddings, which helps the model to filter out hundreds of relevant videos from the original corpus of millions of YouTube videos. This stage enables a preliminary selection via a generalization of matrix factorization. Then, the second component *ranking* network assigns a score to each video based on user and video embedding and outputs the highly personalized recommendation list.

Hypothesis: The YouTube DNN model performs better in recognizing users' long-term intention with a longer sequence length.

Reasoning: The input of the model is a sequence of users' previous behavior, and the task is to predict only the next watched video. As the model averages the embeddings of user watches, a longer sequence can better encode long-term intentions.

3.3.2 GRU4Rec: Recurrent Neural Network (RNN)

GRU4Rec is a RNN-based model for session-based recommendations, where RNN is first introduced in recommendation systems [25]. The initial input is the first item interacted with the user, and each subsequent interaction will give a recommendation that depends on all the previous user behavior. RNN features its ability to model variable-length sequence data along with the hidden state in each unit of the network. The modified GRU4Rec model discounts the input embedding that occurred earlier for the weighted sum, which enables the model to focus more on recent data. Basically, GRU4Rec can be treated as a baseline model for both deep learning models and session-based models in the recommendation.

Hypothesis: The GRU4Rec outperforms the YouTube DNN model in terms of both short-term and long-term user intention.

Reasoning: Compared with YouTube DNN, GRU4Rec generates each recommended item based on all the previous user behavior, reinforcing its long-term memory. In addition, it puts more weight on recent data, allowing a better prediction of the short-term intention.

3.3.3 MIND

MIND is a state-of-the-art model used in the Mobile Tmall App consisting of a multi-interest extractor layer based on the capsule routing mechanism and a label-aware attention layer [5]. It generates multiple representation vectors for each user embedding that correspond to different interests, namely user intention in our work. MIND has proved to achieve higher recommendation accuracy than other existing models.

Hypothesis: MIND outperforms YouTube DNN in both short-term and long-term user intention.

Reasoning: Compared with YouTube DNN, although both of them utilize a deep learning model to generate user embeddings in a large-scale dataset, MIND generate multiple vectors for each user to represent user interests, while the user interest of YouTube DNN is a single vector. Given its advantage in learning multiple interests, we expect it to outperform YouTube DNN in both short-term and long-term user intention.

3.4 ComiRec-SA

ComiRec-SA is a controllable multi-interest self-attentive recommendation model proposed in [4] that trains the model to learn users' multiple intentions from sequential behavior data. We expect ComiRec to outperform other baseline models in our proposed intention-based metric, given its focus on users' multiple interests.

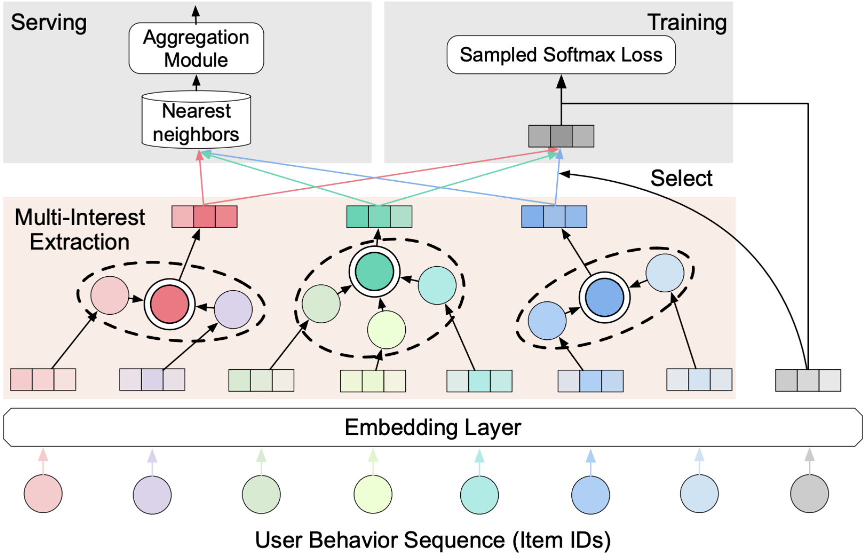

Figure 5: Model Structure of ComiRec [4]

As shown in Figure 5, the ComiRec model can be decomposed into four correlated components: computation of user embeddings, multi-interest extraction, training model based on sampled softmax loss, and the generation of a recommended item set. Among these four, the quality of user embeddings computed from previous user behavior largely determines the performance of the matching stage of recommendation.

The input of the model is a user behavior sequence that contains a sequence of item IDs that the user has interacted with in time order. The item IDs are passed through the embedding layer to generate item embeddings. With these item embeddings, the multi-interest extraction module is able to create multiple interest embeddings for model training and serving. In the stage of model training, the sampled softmax loss is computed with the nearest interest embedding to the target embedding. As for model serving, each interest embedding will independently retrieve top-N nearest items, which will be passed through the aggregation module to get the final output. Here, the aggregation module combines the recommendations from each interest to generate the overall top-N items with a balance between accuracy and diversity.

Hypothesis: ComiRec-SA outperforms all the three models above in both short-term and long-term user intention.

Reasoning: According to the experiment results shown in [4] , ComiRec-SA outperforms MIND in capturing user intention with a balance of accuracy and diversity. Hence, we argue that ComiRec-SA would perform the best in our proposed metric as well.

3.5 Evaluation Metrics

The models designed for recommendation systems can be evaluated from multiple aspects, of which the most two popular ones are accuracy and diversity. It has been shown that there exists a trade-off between accuracy and diversity. If a model solely aims at predicting exactly what the user would click next, when the accuracy has achieved a certain high level, the recommended item set is very likely to reinforce the user's previous interests. Our main goal is to put forward a practical and universal metric to evaluate whether a model can capture users' intentions well or not. The "universal" here means that (i) the metric can be applied to both short-term and long-term intention evaluations, and (ii) it can work well in any typical dataset generated for recommendations, not just the *movieLens* dataset.

3.5.1 Accuracy metrics

Recall at the individual user level can be defined as

$$Recall@N = \frac{1}{\mid \mathcal{U} \mid} \sum_{u \in \mathcal{U}} \frac{\mid \hat{\mathcal{I}}_{u,N} \cap \mathcal{I}_u \mid}{\mid \mathcal{I}_u \mid} \tag{1}$$

where $\hat{\mathcal{I}}_{u,N}$ denotes the set of top-N recommended items for user u and $\mathcal{I}_u$ denotes the testing item set for user u [26].

Hit Rate(HR) is defined as the ratio of sales of a product to the number of total customers who have viewed it [26]. In recommendation systems, if the user selects at least one item from the recommendation lists, we consider it as one "hit". The HR of the whole system is then defined as the count of hits divided by the number of users involved.

$$HR@N = \frac{1}{\mid \mathcal{U} \mid} \sum_{u \in \mathcal{U}} \delta(\mid \hat{\mathcal{I}}_{u,N} \cap \mathcal{I}_u \mid > 0) \tag{2}$$

where $\delta(\cdot)$ denotes the indicator function.

Normalized Discounted Cumulative Gain(NDCG) penalizes the selected item that appear lower in the recommended list by reducing its "gain" [27].

$$NDCG@N = \frac{1}{Z} \frac{1}{\mid \mathcal{U} \mid} \sum_{u \in \mathcal{U}} \sum_{k=1}^{N} \frac{Gain_k}{log_2(k+1)} \tag{3}$$

where Z is the normalization constant, the maximum possible value of $DCG@N$. For simplicity, we can define $Gain_k$ as the indicator function $\delta(\hat{i}_{u,k} \in \mathcal{I}_u)$ where $\hat{i}_{u,k}$ denotes the $k-$th recommended item for the user u.

3.5.2 Diversity metrics

$$Diversity@N = \frac{\sum_{j=1}^{N-1} \sum_{k=j+1}^{N} 1 - Similarity_{j,k}}{N(N-1)/2} \tag{4}$$

where

$$Similarity_{j,k} = \frac{CATE(\hat{i}_{u,j}) \cap CATE(\hat{i}_{u,k})}{CATE(\hat{i}_{u,j}) \cup CATE(\hat{i}_{u,k})} \tag{5}$$

and $CATE(i)$ maps item i to its category. In MovieLens, this function returns a set of genre labels for each given movie.

3.5.3 Intention metrics

$$IntentionRecall@N = \frac{1}{\mid \mathcal{U} \mid} \sum_{u \in \mathcal{U}} \sum_{k=1}^{N} \delta(\hat{i}_{u,k} \in \mathcal{T}_u) \tag{6}$$

where $\mathcal{T}_u$ denotes the set of intentions for user u.

The intention metric can be interpreted as an intention-based *Recall*, in that the only difference is the meaning of the ground truth data, which refers to the test item set in *Recall* and the user intention set in *Intention* respectively. This metric is applicable to variable-length time periods of intentions, allowing us to measure both long-term intention and short-term intentions via the same metric in the following section.

4 Results and Discussion

4.1 Experimentation protocol

The experiment used our newly proposed techniques, user intention extraction and *intention recall* metric, to demonstrate the shortage of current metrics for model evaluation in recommendation systems and further show the rationality of the new metric. In all of our experiments, the recommendation models generate 10 items as output. However, it is also vital to adjust parameters to better evaluate their performance in capturing long-term and short-term intentions. We focus on two parameters: (1) *session length*, which is related to the definition of short-term and long-term, and (2) *sequence length*, which is the input length for the sequence model. The two parameters correspond to intention definition and model choice respectively. As the detection of short-term intention only depends on the latest session, the ideal definition of a short term can be discovered by modifying the value of *session length* to identify which value enables the best score in short-term intention evaluation. As for the model choice, we can find the best model in each given *session length* by adjusting the value of *sequence length.* In the experiment, the metric *intention recall* is used as a crucial measurement. The higher the *intention recall* is, the better a model's performance is. If a model consistently outperforms regardless of parameter changes, it has a strong capacity to capture user intents. In our experiments, we tested each model with three *sequence length*: 5, 100 and 256, with four *session length*: 1, 5, 10, 20. The experiment was conducted in Google Colab via the Tensorflow framework. The models used in the experiment are DNN, GRU4REC, MIND, and ComiRec-SA [1].

4.2 Experiment Results

4.2.1 Misalignment with accuracy and diversity

Table 2 compares the accuracy of four types of models and models with distinct parameters when the number of recommendation items is 10, i.e., N=10, and *seq* denotes the input length for sequential models. We can see that MIND and ComiRec both achieve the best accuracy at seq = 5. For DNN and GRU4REC, the best accuracy occurs at seq = 100 and seq = 256 respectively. Theoretically, there is a trade-off between accuracy and diversity. The higher performance on accuracy will result in lower performance in diversity. But for MIND and ComiRec, they achieve the best accuracy and diversity at the same time, which is also claimed in [4]. This inspires

[1] Code is available at https://github.com/THUDM/ComiRec

Model	Recall@10	NDCG@10	Hit Rate@10	Diversity
DNN (seq = 5)	0.047542	0.170117	0.360656	**0.392405**
DNN (seq = 100)	**0.067284**	**0.248777**	**0.442623**	0.391979
DNN (seq = 256)	0.043127	0.204712	0.426230	0.321311
GRU4Rec (seq= 5)	0.061335	0.229794	**0.426230**	**0.439367**
GRU4Rec (seq = 100)	0.047386	0.210937	0.409836	0.411223
GRU4Rec (seq = 256)	**0.062572**	**0.252062**	0.393443	0.402351
MIND (seq= 5)	**0.061727**	0.212409	**0.409836**	**0.405181**
MIND (seq = 100)	0.048894	0.199501	0.393443	0.377984
MIND (seq = 256)	0.047878	**0.223693**	0.393443	0.396922
ComiRec-SA (seq = 5)	**0.030747**	0.170921	**0.409836**	**0.442702**
ComiRec-SA (seq = 100)	0.029017	**0.176130**	0.344262	0.376589
ComiRec-SA (seq = 256)	0.027940	0.126863	0.311475	0.406481

Table 2: Accuracy metrics across all tested models.

us to derive a new metric for model evaluation from another aspect. With our *intention recall* metric, the results reveal misalignment with accuracy and diversity. To be specific, as shown in Table 3 and Table 4, both of the models with the highest scores in our intention-aware metric are trained at seq = 256, which is inconsistent with their performance in accuracy and diversity. These findings show that the currently existing metrics are not sufficient for model evaluation. More importantly, our metric is able to provide new insights regarding model evaluation.

4.2.2 Performance on user intention

Table 3 and Table 4 compare the ability of different models to capture long-term and short-term user intention. *LT* is short for Long-Term Intention and *ST* is short for Short-Term Intention. As for the notation in the parenthesis, *s* stands for session and the number following denotes the length, i.e., s5 means each session includes 5 rating records, which is the length of short-term session as well. From these two tables, *intention recall* are under the influence of multiple factors and needs to be further explored as follows.

Variant: Session Length. Figure 6 depicts the ability of each model to capture the long-term and short-term user intention with multiple values of session length and the fixed input sequence length at 256. The first discovery is that as the *session length* increases, the *intention recall* would decrease accordingly. This is understandable. Because the extraction of user intents renders the long-term and short-term intentions mutually exclusive, long-term sessions do not contain the most recent session, namely the short-term session. As the length of the session increases from 1,

Model	LTI (s1)	STI (s1)	LTI (s5)	STI (s5)	Diversity
DNN (seq = 5)	0.496721	0.342623	0.496721	0.381967	**0.392405**
DNN (seq = 100)	0.526230	**0.380328**	0.526230	**0.406557**	0.391979
DNN (seq = 256)	**0.560656**	0.344262	**0.560656**	0.381967	0.321311
GRU4Rec (seq = 5)	**0.559016**	**0.396721**	**0.559016**	0.395082	**0.439367**
GRU4Rec (seq = 100)	0.552459	0.386885	0.552459	0.403279	0.411223
GRU4Rec (seq = 256)	0.544262	0.380328	0.532787	**0.431148**	0.402351
MIND (seq = 5)	0.568852	0.373770	**0.568852**	0.380328	**0.405181**
MIND (seq = 100)	0.549180	0.383607	0.549180	0.391803	0.377984
MIND (seq = 256)	**0.575410**	**0.398361**	0.531148	**0.449180**	0.396922
ComiRec (seq = 5)	0.545902	0.367213	0.545902	0.411475	**0.442702**
ComiRec (seq = 100)	0.554098	0.426230	0.554098	0.444262	0.376589
ComiRec (seq = 256)	**0.593443**	**0.432787**	**0.567213**	**0.455738**	0.406481

Table 3: Intention and diversity metrics for top10 recommendations across all tested models. Intention metrics use session length 1 and 5.

Model	LTI (s10)	STI (s10)	LTI (s20)	STI (s20)	Diversity
DNN (seq = 5)	0.370492	0.303279	0.272131	0.213115	**0.392405**
DNN (seq = 100)	0.403279	**0.326230**	0.308197	0.232787	0.391979
DNN (seq = 256)	**0.422951**	0.321311	**0.318033**	**0.237705**	0.321311
GRU4Rec (seq = 5)	**0.436066**	**0.327869**	**0.313115**	**0.249084**	**0.439367**
GRU4Rec (seq = 100)	0.413115	0.314754	0.290164	0.226230	0.411223
GRU4Rec (seq = 256)	0.422951	0.318033	0.308197	0.237705	0.402351
MIND (seq = 5)	**0.444262**	0.332787	**0.319672**	0.244262	**0.405181**
MIND (seq = 100)	0.432787	0.318033	0.309836	0.239344	0.377984
MIND (seq = 256)	0.426230	**0.342623**	**0.319672**	**0.267213**	0.396922
ComiRec (seq = 5)	**0.427869**	0.344262	0.318033	0.260656	**0.442702**
ComiRec (seq = 100)	0.421311	0.342623	0.311475	0.254098	0.376589
ComiRec (seq = 256)	0.426557	**0.358033**	**0.329836**	**0.279180**	0.406481

Table 4: Intention and diversity metrics for top10 recommendations across all tested models. Intention metrics use session length 10 and 20.

the length of the long-term reduces, and at the same time the length of the short-term increases. The boundary of long-term and short-term gradually become vague. Therefore, the extracted user intention can not be a ground truth any more as session length goes up. In this experiment, the best long-term intention is reached when the session length is 1, while the best short-term performance is produced when the session length is 5.

Figure 7 compares the *intention recall* among 4 models when each model reaches its maximum intention. The session length of the data points are different here. ComiRec surpasses the other three models in terms of both long-term and short-term intention. In these two areas,

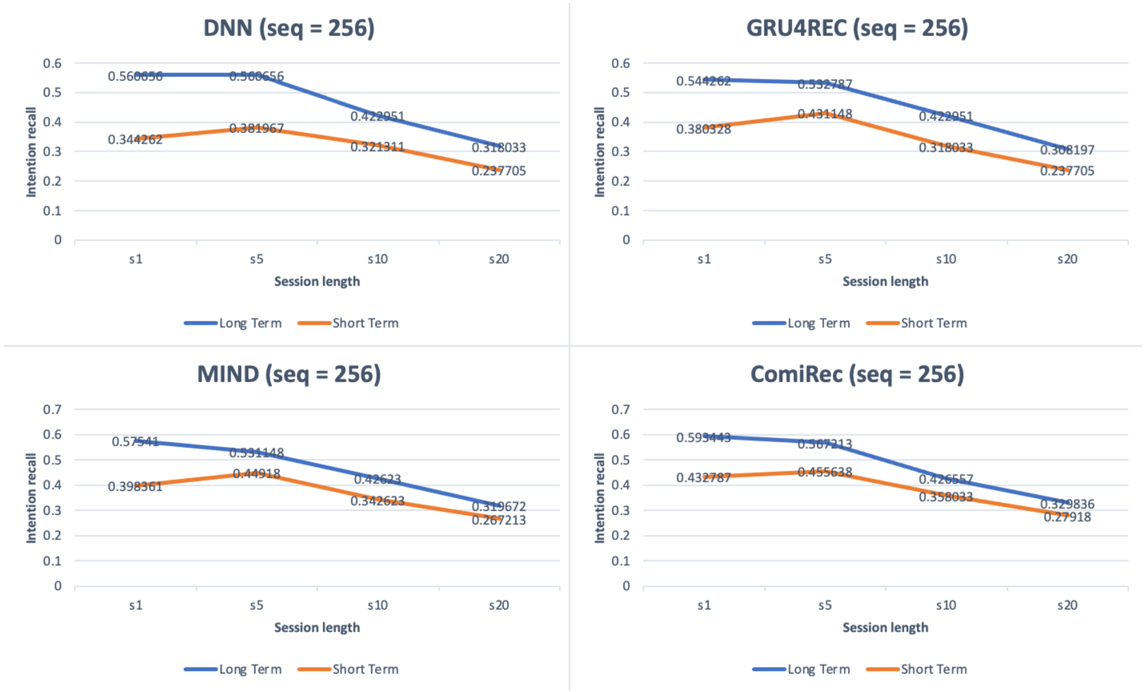

Figure 6: the variation of intention capture of the 4 models(seq = 256)

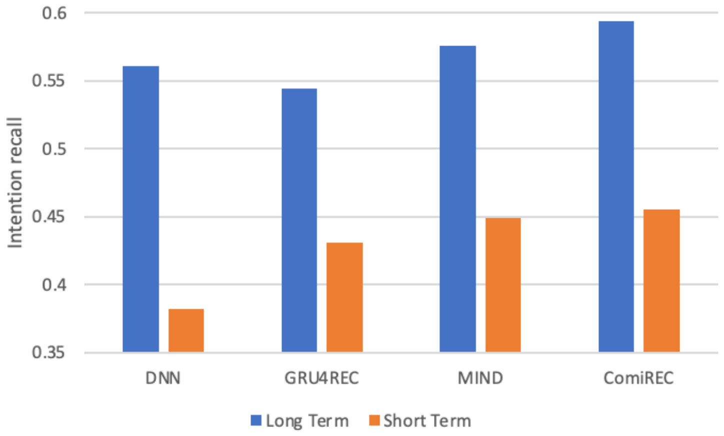

Figure 7: best intention recall comparison among 4 models (seq = 256)

MIND outperforms DNN and GRU4REC. DNN is the poorest at capturing short-term intentions, while GRU4REC is the worst at capturing long-term intentions. This result is in line with the experiment's expectations. This also further proves the rationality of the new metric *intentional recall* and the extraction of user intention.

Variant: Sequence Length. The performance on intention of four models with varied sequence lengths is ranked in table Table 5. As a result, the total number of candidates in the ranking is 9. The majority of models in the top 1 to 4 rankings use sequence length 256. For these models,

	Top1	Top2	Top3	Top4
LTI(s1)	ComiRec (seq = 256)	MIND (seq = 256)	DNN (seq = 256)	GRU4Rec (seq = 5)
LTI(s5)	MIND (seq = 5)	ComiRec (seq = 256)	DNN (seq = 256)	GRU4Rec (seq = 5)
LTI(s10)	MIND (seq = 5)	GRU4Rec (seq = 5)	ComiRec (seq = 5)	DNN (seq = 256)
LTI(s20)	ComiRec (seq = 256)	MIND (seq = 256)/MIND (seq = 5)	DNN (seq = 256)	GRU4Rec (seq = 5)
STI(s1)	ComiRec (seq = 256)	MIND (seq = 256)	GRU4Rec (seq = 5)	DNN (seq = 100)
STI(s5)	ComiRec (seq = 256)	MIND (seq = 256)	GRU4Rec (seq = 256)	DNN (seq = 100)
STI(s10)	ComiRec (seq = 256)	MIND (seq = 256)	GRU4Rec (seq = 5)	DNN (seq = 100)
STI(s20)	ComiRec (seq = 100)	MIND (seq = 256)	GRU4Rec (seq = 5)	DNN (seq = 256)

Table 5: Different Model Best Results Rank

however, the best accuracy is not obtained at seq 256 as shown in Table 2. In Table 2, when the sequence length is 5, both the ComiRec and MIND models achieve their best accuracy. The sequence length is the input length for sequential models. The longer the sequence length is, the better it would be able to collect user intents. But as the sequence length decreases, the higher accuracy the model will reach. The importance of the new metric *intention recall* is further demonstrated by the trade-off between accuracy and intention.

This table also proves that ComiRec and MIND outperforms the other two models when detecting user intention, which aligns with the expectation.

5 Discussion

5.1 Limitations in data collection

The features of the dataset itself largely affect the model performance. Despite its wide application in studies of recommendation systems, the available data attributes of MovieLens datasets limit models' learning abilities towards user intention. The user-item interactions are only represented by users' rating histories, which can only partially account for user intention. We have realized that the majority of ratings occurred on the user's first day of using the system, which poses a challenge for intention capturing on that special day. Intuitively, the rated movies of the first day can somehow imply users' long-term intention in that they can be seen as a summary of all past viewing records. To deal with it, our method treats each rating history as sequential data due to the sparsity of the dataset, and the temporal information only plays a role in data discovery as a preliminary stage. However, if we can get access to users' viewing histories as well, then user intention would be captured more accurately.

Another limitation lies in the artificial extraction of user intention. In our study, the most popular item categories serve as user intention which are assumed to be the ground truth, but they are not. All of our experiments are conducted in the offline setting, and however, if users can indicate their intentions through novel designs of user interaction, for example, the system asks for user inputs after they've browsed several items. Such mechanisms to elicit user intent inputs will increase the reliability of intention generation for our evaluations.

Furthermore, the definitions can vary a lot depending on different settings. For instance, as for recommendations on social media platforms, users may welcome contents that go beyond their expectations to break down their filter bubbles. From that standpoint, user intention can be defined to distinguish the dedicated preferences from the purpose of deliberately receiving new information. Last but most importantly, the evaluation of metrics requires a bunch of experiments on various models. The challenge is to identify the influence of the model and the metric. Any novel metric can be justified only if it performs consistently on different kinds of models and the results can be reasonably interpreted.

5.2 Future work on metric design

Evaluating a new metric is a tricky problem, and several questions still remain to be answered. The intention-based recall metric only utilizes a binary indicator function to measure the extent to which the output list aligns with the ground truth in terms of user intention. That is, how the two sets of item categories match with each other. But in fact, we can measure the alignment in more dimensions, such as the position of each recommended item (similar to the logic of NDCG), the proportion of each category in the recommendation list, and the decaying weight on older sessions than the recent ones, etc.

Given the inconsistent performances of some models in short-term and long-term intention capturing, an intention-aware loss function in the model training stage may help a lot. In our study, we focus exclusively on the metric design and the verification tests on its rationality. As for models, we consider them as "black boxes" to serve as our experimental subject. That is, we made no modifications to the training target function or the model structure. In a similar manner to the intention recall definition, future studies may modify the traditional loss functions to reflect the influence of user intention as well. The improvements in models can be a strong supplementary part of perfecting the metric design.

6 Conclusion

The main contributions of this research can be divided into two parts: the detection mechanism of user intention and a novel intention-aware recall metric. Unlike accuracy metrics, there's no universal measurement of model performance in terms of user intention, especially due to the high flexibility in its definition. Inspired by the increasing focus on user intention in recommendation models, we offered a new aspect of model evaluation accordingly via a two-step approach: (1) user intention extraction and (2) empirical experiments on baseline models toward our new proposed metric *intention recall*. The first part can be skipped if the ground-truth intentions are available in online settings. Otherwise, our statistical analysis may help to extract intentions artificially. Our experiments show that *intention recall* can provide new insights in addition to accuracy and diversity. Moreover, the conclusion drawn from our metric matches well with the quantitative result in [4], demonstrating our new metric's usefulness and applicability.

References

[1] X. Liu, Y. Liu, K. Aberer, and C. Miao, "Personalized point-of-interest recommendation by mining users' preference transition," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 733–738.

[2] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert Systems with Applications*, vol. 69, pp. 29–39, 2017.

[3] C. Zhou, J. Bai, J. Song, X. Liu, Z. Zhao, X. Chen, and J. Gao, "Atrank: An attention-based user behavior modeling framework for recommendation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[4] Y. Cen, J. Zhang, X. Zou, C. Zhou, H. Yang, and J. Tang, "Controllable multi-interest framework for recommendation," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2942–2951.

[5] C. Li, Z. Liu, M. Wu, Y. Xu, H. Zhao, P. Huang, G. Kang, Q. Chen, W. Li, and D. L. Lee, "Multi-interest network with dynamic routing for recommendation at tmall," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 2615–2623.

[6] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *2008 Eighth IEEE international conference on data mining.* Ieee, 2008, pp. 263–272.

[7] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 285–295.

[8] J. Wang, A. P. De Vries, and M. J. Reinders, "Unifying user-based and item-based collaborative filtering approaches by similarity fusion," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 501–508.

[9] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," in *International conference on algorithmic applications in management.* Springer, 2008, pp. 337–348.

[10] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE International Conference on Data Mining (ICDM).* IEEE, 2018, pp. 197–206.

[11] C.-Y. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing, "Recurrent recommender networks," in *Proceedings of the tenth ACM international conference on web search and data mining*, 2017, pp. 495–503.

[12] D. Aarno and D. Kragic, "Motion intention recognition in robot assisted applications," *Robotics and Autonomous Systems*, vol. 56, no. 8, pp. 692–705, 2008.

[13] G. Khodabandelou, C. Hug, R. Deneckere, and C. Salinesi, "Supervised intentional process models discovery using hidden markov models," in *IEEE 7th International Conference on Research Challenges in Information Science (RCIS).* IEEE, 2013, pp. 1–11.

[14] K. Yao, G. Zweig, and B. Peng, "Attention with intention for a neural network conversation model," *arXiv preprint arXiv:1510.08565*, 2015.

[15] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[16] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[17] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu, "A theoretical analysis of ndcg ranking measures," in *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, vol. 8. Citeseer, 2013, p. 6.

[18] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: evaluating recommender systems by coverage and serendipity," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 257–260.

[19] E. Amigó, D. Spina, and J. Carrillo-de Albornoz, "An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 625–634.

[20] M. Slaney and W. White, "Measuring playlist diversity for recommendation systems," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, 2006, pp. 77–82.

[21] P. Castells, N. Hurley, and S. Vargas, "Novelty and diversity in recommender systems," in *Recommender systems handbook*. Springer, 2022, pp. 603–646.

[22] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A comparative analysis of cascade measures for novelty and diversity," in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 75–84.

[23] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, pp. 1–19, 2015.

[24] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.

[25] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.

[26] G. Karypis, "Evaluation of item-based top-n recommendation algorithms," in *Proceedings of the tenth international conference on Information and knowledge management*, 2001, pp. 247–254.

[27] K. Järvelin and J. Kekäläinen, "Ir evaluation methods for retrieving highly relevant documents," in *ACM SIGIR Forum*, vol. 51, no. 2. ACM New York, NY, USA, 2017, pp. 243–250.