School of Informatics

Informatics Project Proposal
Recommender Systems: looking further into the future

S2406121
April 2023

Abstract

Recommender systems have already become a popular topic with the development of artificial intelligence, which could provide a personalized user experience increasing customer satisfaction. However, traditional recommender systems are mostly based on the next-item prediction, which only focuses on the next single action of the user, and cannot predict the whole scope of user interests. In this project, we introduce the multi-day window predictor to the recommender system that aims to predict all items that the user may be interested in within a time window. In theory, this novel architecture could lead to better recommendation results, reduce the cost of deployments, and be more robust against short-term variations.

Date: Tuesday 11th April, 2023

Tutor: Aurora Constantin
Supervisor: David Wardrope, Timos Korres, Michael la Grange

1 Motivation

The personalized experience is one of the most attractive user experiences in different social media, which has already attracted a lot of attention recently. Recommendation system has become a popular topic[1], which helps users to find their interests and new items based on user's previous preferences. The performance of a recommendation system can directly influence user satisfaction in e-commerce, streaming, and social media platform. Currently, Many different kinds of approaches to improve the performance of recommendation systems have been proposed[2] including traditional models and deep learning models. In recent years, many sequential models, such as recurrent neural network (RNN), have been proven to be very useful in processing sequential data, which is also widely used in recommendation system[3, 4, 5, 6] because users' previous action can also be seen as an input sequence. Due to the great improvement brought by the sequential neural recommender, the performance of the current recommendation system has been shown very effective. However, most existing methods always focus on next-item prediction, which aims to correctly recommend the single next item that the user will interact with, and this method may not capture the sufficient scope of user preferences and behaviour. Furthermore, focusing on predicting the next single item can be easily influenced by short-term variations, which also require more frequent deployments. Therefore, this project aims to further improve the performance of recommender systems by implementing a window-based sequential recommendation system, which can correctly identify all items the users may interact with within a time window.

1.1 Problem Statement

This project aims to improve the performance of the current recommendation system by proposing a new window-based sequential recommender and examining the effectiveness of this architecture.

We can further decompose the project into three subtasks. In the first subtask, we will implement a sequential neural recommendation system for next-item prediction incorporating many state-of-the-art approaches to ensure the performance of the model. In the second subtask, we will extend the constructed model to a window-based recommendation system, which aims to predict all items in a given time window. The final subtask will focus on evaluating and analysing the performance of the window-based model with next-item prediction models including different existing models[3, 4, 6] and also some very simple baseline models[2] to prove the effectiveness of the window-based predictor. Overall, this project will explore how to build a window-based sequential recommendation system in detail and whether this window architecture could outperform the next-purchase prediction approach that is most commonly used. Following these steps, we will demonstrate the performance of the proposed model and whether the window-based sequential recommender system could lead to improving the standard approach of predicting users' single next purchase.

1.2 Research Hypothesis and Objectives

The aim of the project is to examine whether predicting the sequence of a customer's future purchases is a better learning task for recommender systems than the standard approach of predicting their single next purchase.

To achieve this goal, we will split the whole project into three objectives including implementation, extension, and evaluation. we will first implement a sequential recommendation system for next-item prediction, which is expected to be more effective than some of the state-of-the-art approaches, then, extend the next-item prediction model to a multi-window predictor and analyse the performance of the multi-window predictor. Finally, we will evaluate the model performance compared with different existing models.

1.3 Timeliness and Novelty

The research is proposed very timely because it can utilize and incorporate many existing useful methods to improve the performance of the sequential recommendation system including scheduled sampling[7] in 2016 and contrastive learning with robust augmentation[8] in 2021. Moreover, the personalized user experience has become more and more popular with the development of the current technology, so it is necessary to improve the performance of the recommendation system to fit the users' requirements. The window-based recommender system PinnerFormer is firstly proposed by Pancha et al.[9] in 2022 to better predict user interest in Pinterest, which shows an improvement compared with the SASRec[4], a transformer-based next-item prediction system. Therefore, it is meaningful to examine whether this time window architecture could actually lead to better performance of the sequential recommendation system.

1.4 Significance

ChatGPT of OpenAI almost attracts worldwide attention to the current artificial intelligence, which also leads to a boom in AI development. Therefore, an improvement in every field of artificial intelligence is very significant including recommender systems. In recent recommendation systems, the next-item prediction is still the most widely used method, but this method is less robust and hard to capture the large scope of user interests because it only focuses on the next single action. Therefore, it is essential to propose some new approaches to improve the model performance and robustness. This project will introduce a new window prediction approach and examine the effectiveness of the approach, which could potentially improve the current prediction method of sequential recommender systems. If the window prediction has been shown more effective compared with the next-item prediction approach, it will help further development in the field of recommender systems, which may change the research direction from traditional next-item prediction to time window prediction. Furthermore, this project can also incorporate some currently proposed methods[7, 8, 10] for sequential recommendation systems, which can potentially help to achieve better performance when predicting based on user preference.

1.5 Feasibility

The proposed plan of the project consists of implementing a next-item prediction system, extending the time-window architecture and the evaluation. We have found several alternative sequential recommenders in the relevant topic. Meanwhile, the code implementation of many typical models for sequential recommendation systems[3, 4, 6] could be found in GitHub. The implementation and extension part of the project can be carried out by modifying the existing models including changing the loss function and changing different hyperparameters. For the final evaluation task, we need to compute the performance of different models, which could need a lot of time and computational cost. To address this problem, we will access informatics GPU (MLP cluster), which could provide the sufficient computational ability to address the issue. By separating the whole project into several small tasks, the project could be completed within the

10-week time frame. These steps show the whole progress could be completed on time before the due time.

1.6 Beneficiaries

The research project will benefit all users of the system and may improve the development of recommender systems from next-item prediction to window-based prediction. As described above, the next-item prediction approach is still the most widely used approach, which may not be effective to capture the entire scope of users' interest, if the proposed next-window prediction approach could be proven to improve the model performance compared with the traditional method, it will significantly influence the field. The performance of recommendation systems in different domains such as e-commerce and movie recommendation may be improved, which will facilitate the development of sequential recommendation systems and also result in higher user satisfaction.

2 Background and Related Work

The development of recommendation systems moves fast, many existing approaches have been proposed and shown effectiveness to some extent. In this section, we introduce different approaches to recommendation systems including traditional models and deep learning models in this section. The strengths and weaknesses of each method are discussed.

Beginning with traditional recommendation systems, traditional recommendation systems often use the first-order Markov Chain for capturing the short-term dependency within a behaviour sequence, which only focuses on the most recent previous single action of the user. Most of the conventional approaches ignore the influence of the sequential dependencies among the user's interactions, which results in an inaccurate recommendation. However, It is also deserved to analyse the performance of conventional recommendation approaches, which can be used as useful baselines in the project.

PopRec is the simplest model, which can be easily implemented by always recommending the top popular items to the users. However, this method is very commonly used as a baseline in many famous recommendation systems [4, 3].

K-nearest neighbors(KNN) is also used for sequential recommendation. Davidson et al. [11] use the KNN method to produce a YouTube video recommendation system, which has been demonstrated to have an improvement compared with some very basic recommenders. Item-based KNN method, which only considers the last behaviour in a sequence and uses the most recent behaviour to calculate the similarity, is a highly explainable approach. However, the KNN recommendation system cannot capture the sequential dependency in the input sequence, which lead to a bad recommendation.

Matrix Factorization (MF) learns the latent factors for items and users, which is still commonly used as a layer in the current neural recommender. In 2009, Rendle et al.[12] proposed Bayesian Personalized Ranking (BPR), an MF-based recommendation method. BPR learns the user embedding and item embedding from the user-item interaction matrix and uses the specific user embedding to calculate the score of all items in the item embeddings. The difference between this method with standard MF is the optimization objective, BPR aims to rank the relevant positive interaction higher than the negative interaction whereas the standard MF method only considers minimizing the distance between positive items and the user. The implementation based on Netflix Prize shows an improvement of BPR compared with the standard matrix factorization.

Markov chain is commonly applied to build a sequential recommendation system. Especially, many basic recommendation systems are based on the first-order Markov chains, which only pay attention to the last interacted item. Factorized Markov Chain (FMC) uses the matrix factorization method based on the first-order Markov chain to generate the recommendation. Specifically, this approach factorizes the item-transition matrix into two low-dimensional item embeddings, then, use the embedding of the last visited item to calculate a score of all other items. In addition, Rendle et al.[13] in 2010 proposed a more advanced Factorizing Personalized Markov Chains(FPMC) for sequential recommendation extending from the factorized Markov chains. FPMC factorizes the user-item matrix and item-item transition matrix to get the user embedding and the item embedding, the first factorization aims to obtain the long-term user preference and the second factorization aims to capture the short-term dependency as the same as FMC. This method shows an obvious improvement compared with the normal FMC and standard matrix factorization methods. Although all these first order Markov chains-based approaches can generate highly explainable recommendations but due to the limitation of first order Markov chains, they are hard to capture the long-term dependency in the behaviour sequence effectively.

In recent years, with the development of machine learning algorithms, deep learning methods especially some sequential modelling methods including RNNs and transformers are demonstrated to be more effective when processing sequential data. Therefore, current research on sequential recommendation systems focuses on the improvement of sequential modelling methods. Some famous approaches to sequential recommendation will be introduced below.

GRU4Rec is the first model applying recurrent neural networks (RNN) on the recommender systems proposed by Hidasi et al.[3] in 2016, which is a significant milestone in the recommender systems. GRU4Rec use the gated recurrent unit (GRU) architecture to handle the user behaviour sequence to better capture the long dependencies in the interaction sequence, which significantly outperforms the traditional machine learning methods, especially handling long sequence. This model processes the historical sequence of the user and uses the final state of the GRU to represent the user preference, the final recommendation represented by calculating the score of the user preference and all item embeddings. The main advantage of this method is that it can better handle long sequences of user interactions, which is significant because users' history may have a long sequence. GRU4Rec has been shown to outperform other state-of-the-art recommendation algorithms in terms of recommendation accuracy. Although GRU4Rec improve the progress of handling long sequence it is still hard to handle very long sequence and it also lacks the interpretability of its recommendation, the same as many other neural networks, it is hard to know why the model predicts a specific recommendation list. Moreover, GRU4Rec Could still suffer from cold-start problems because GRU4Rec relies on users' historical sequence to make a prediction. If the user has no previous interaction or does not have sufficient historical actions, the model will not perform well.

As the self-attention mechanism is proposed by Vaswani[14] in 2017, the transformer becomes a very popular architecture in sequential modelling methods. **SASRec** is the first transformer-based recommender system proposed by Kang et al.[4] shows a better performance in handling very long sequences and enhances the interpretability of the recommender systems. The self-attention mechanism with causal masking computes the user representation by allowing the system to capture the most relevant items in the historical actions of the user regardless of their temporal distance, then, computes the weighted sum of different item embeddings in the previous sequence to build a more reliable user representation. In this case, SASRec is capable to handle an unlimited length of the sequence. Meanwhile, the self-attention mechanism calculates the weight of different items in the input sequence separately, which gives a clear view of the importance of different items. Therefore, SASRec also has better interpretability. However, SASRec has more learnable parameters, which could cause a higher computational cost. Additionally, calculating the user representation of the current state involves all previous items in the sequence, which could cause a higher latency in the online server[10].

PinnerFormer proposed by Pancha et al.[9] in 2022 introduce the window-based prediction method, and shows an improvement in recommender systems. Similar to SASRec, the model handles the sequence by the same transformer architecture to calculate a sequence of user embeddings. Different from the SASRec, only the final user representation is used to generate the recommendation list, PinnerFormer utilizes the intermediate user representation, which aims to predict a randomly selected positive action from the set of all positive actions within the time window, this dense all action prediction also improve the signal pass in the network. The time window architecture and dense all action prediction introduced by Pinnerformer improve the performance compared with SASRec. Whereas, the dense all action prediction considers the intermediate user representation, which could increase the computational cost of the system.

3 Programme and Methodology

This project proposes a novel multi-day window prediction method in recommender systems to improve the performance of sequential recommendation. Excluding dissertation writing, this project can be divided into 4 work packages (WPs). First WP focus on data handling, second on building a next-item prediction sequential recommender, third on extending the constructed next-item prediction model to a multi-day window prediction model and the final WP consider the experiment results and results analysis.

Data handling: To build a more robust system, we need to find a sufficient number of training data, MovieLens or Retailrocket datasets could be the potential datasets for our project. Due to the specific requirements of our task, we need to handle the dataset in some different ways for next-item prediction and Multi-day window prediction. In Both methods, to make the training process effective the length of the input sequence should at least reach a pre-defined threshold. For instance, we should discard users and items with fewer than 5 related actions in next-item prediction while discarding behaviour sequences with fewer than 10 actions in multi-day window prediction to ensure the window architecture has some influence. Additionally, different splitting strategies [15] could be explored and selected for our task. Standard SASRec[4] chooses the item-based split method while other models may prefer the user-based split strategy. In our project, we should choose an appropriate data-splitting strategy to achieve better performance. Robust augmentation[8] proposed in 2021 introduce many data augmentation approaches to handle the data, which could also be a potential way to address some problems of the dataset.

Next-item prediction model: After handling the dataset, we could obtain suitable training, dev, and test datasets for next-item prediction and multi-day window prediction. The implementation part will start with implementing an existing state-of-the-art model for the next-item prediction using RNNs and transformers that can be used to train on our dataset. Python code implementations of several typical sequential recommenders[3, 4] are shared in GitHub. Therefore, the main task of this work package focuses on implementing an existing model and doing some modifications to potentially improve the model performance on our dataset. After implementing the sequential recommender, evaluate the model performance on different metrics[2] such as precision@k and Recall@k. We should also try some potential improvements to demonstrate whether these approaches could achieve better performance in our model. For instance, scheduled sampling[7] proposed in 2015 could mitigate the discrepancy between the training and inference, it introduces a sampling strategy in which the model will not always choose the target input as its input but is more likely to choose its own prediction as the training process. Moreover, Pi et al.[10] proposed a User Interest Center (UIC) architecture in 2019 incorporating an incremental learning method, which stores fixed-length memory tensor, to address the high latency and storage problem caused by the self-attentive mechanism for realtime inference.

Multi-Day window prediction model: After implementing a next-item prediction sequential recommender, idealy we also have the useful training, dev, test dataset in the first WP for multi-day window predictor. The third work package focus on extending the next-item prediction method to the multi-day window predictor. The main difference between the two predictions mainly focuses on the different loss functions and different data splitting strategies. Due to the related work of PinnerFormer[9] we have already have a basic idea of how to apply the time window to the multi-window predictor. PinnerFormer provides a dense all action loss, which allows the model to better capture a user's longer-term interests rather than only predicting a single next action, this loss function could also be implemented in our multi-day predictor to enhance the model performance. Multi-day window predictor aims to predict all items within a time window rather than the single next item so the training, dev, and test dataset could be very different from the previous next-item prediction model. Therefore, some modifications need to be implemented to adjust the model can be trained on the new dataset.

Further Experiments and Analysis: To further evaluate the model performance of the multi-day window predictor, we will analyse the results of the experiment and try different time windows to further demonstrate the performance of the novel predictor. The final work package will figure out whether the multi-day window prediction could improve the current next-item prediction based on our dataset. Moreover, we can also experiment with different deployment cycles and temporal likelihood weighting to optimize the model performance. Overall, in the last WP, we should try to optimize the model performance as possible and evaluate the model performance with other baseline models.

Limitation: Although we will compare many different baseline models with the multi-day window predictor, the generalizability of the multi-window predictor may be limited by the datasets used because we will only focus on a specific dataset to carry out the experiment. In addition, the computational cost of the multi-day window predictor could significantly increase compared with the next-item prediction method, which may not be afforded by some individuals. Lastly, we will inherit the previous hyperparameter settings of the existing model, but other hyperparameter settings may achieve better performance for multi-day window recommenders.

3.1 Risk Assessment

Risk Name	Likelihood	Severity	Control Measures
Insufficient Data	Low	High	Data augmentation
Computational Resource Constraints	Low	High	Apply for high performance computing
Workload Exceeds Time Limit	Low	High	Develop a detailed project plan
Model Complexity	Medium	Medium	Review the model's complexity and select a suitable model architecture

As shown in table 3.1 above, four risks are identified with different degrees of likelihood and severity. Firstly, sufficient training data is essential for training a deep learning sequential model, if our dataset cannot provide enough data after preprocessing and filtering this will cause a problem in our project, which make the model cannot train sufficiently and suffers from the underfitting problem. Therefore, the severity of this problem is high, but these dataset has already been used multiple times in this field so we think the likelihood should be low. If the situation still occurs, data augmentation could be a useful approach to address the problem. Computational Resource constraints could also be a problem when training a large number of data in a complex deep learning model, but we already have permission to access the MLP cluster and have experience using the GPU, so the likelihood for this problem is low, but if MLP cluster cannot mitigate the problem, it will have a high severity. In this case, we can apply for permission for high performance computing in Bayes Center. The likelihood of workload exceeds time limit is low because we already have a detailed project plan but the severity is high. Lastly, the model complexity may be not very suitable for our project such as BERT4Rec[6], so we need to review the complexity of different models and select a suitable model architecture to start our project, the likelihood and severity are both medium in this problem.

3.2 Ethics

There are no exposed ethical issues with this project using the publicly available dataset. Two potential datasets are both widely used for the development of recommendation systems. The only thing that may cause the bias is the original bias in the training data. For instance, if the training dataset includes more males than females, the sequential recommender may perform better to recommend items for males.

4 Evaluation

Following related work [2, 4, 9], three different automatic evaluation metrics will be performed after the implementation and experiment of each sequential recommendation including Hit Ratio (HR), recall@k and NDCG@k, which are all commonly used evaluation metrics for evaluating the performance of a ranked list. The evaluation result of the test set represents the final model performance in our project. Hit Ratio is a binary matrix representing whether the recommendation list generated by the predictor contains at least one ground truth item. A higher HR value indicates the model is more successful to identify at least one relevant item for the user. Recall@k cut off the recommendation list at position k and then calculates the recall in the top-k recommendation list. The value represents how well the recommender system could identify relevant items from the entire set of relevant items, and the higher value indicates the model is better at identifying the relevant items in the recommendation list. NDCG@k also cutoffs the recommendation list at the position, and compare the real recommendation list with the idea recommendation list. The value represents how close the real recommendation is to the perfect recommendation, and the higher value indicates the high quality of the recommendation system.

5 Expected Outcomes

We expect the multi-day window predictor could implement successfully by extending from the existing sequential recommendation and explore the potential benefits of multi-day window prediction in recommender systems compared to the standard next-item prediction approach. The performance of multi-day window prediction is expected to be better than the next-item prediction method, which indicates that the multi-day window architecture may provide a more accurate recommendation result to users increasing user satisfaction with extending applications. If the model can perform as expected it may contribute to a wide usage of this method in recommendation systems instead of next-item prediction, which is a novel approach in this field. If the multi-day window predictor can achieve higher performance, it will fix some drawbacks of the traditional next-item prediction, which only focuses on the single next user action and may be easily influenced by short-term variations. If the model can successfully extract the temporal likelihood, it could also extend understanding of the temporal aspects of user behaviour and preferences in recommender systems.

6 Research Plan, Milestones and Deliverables

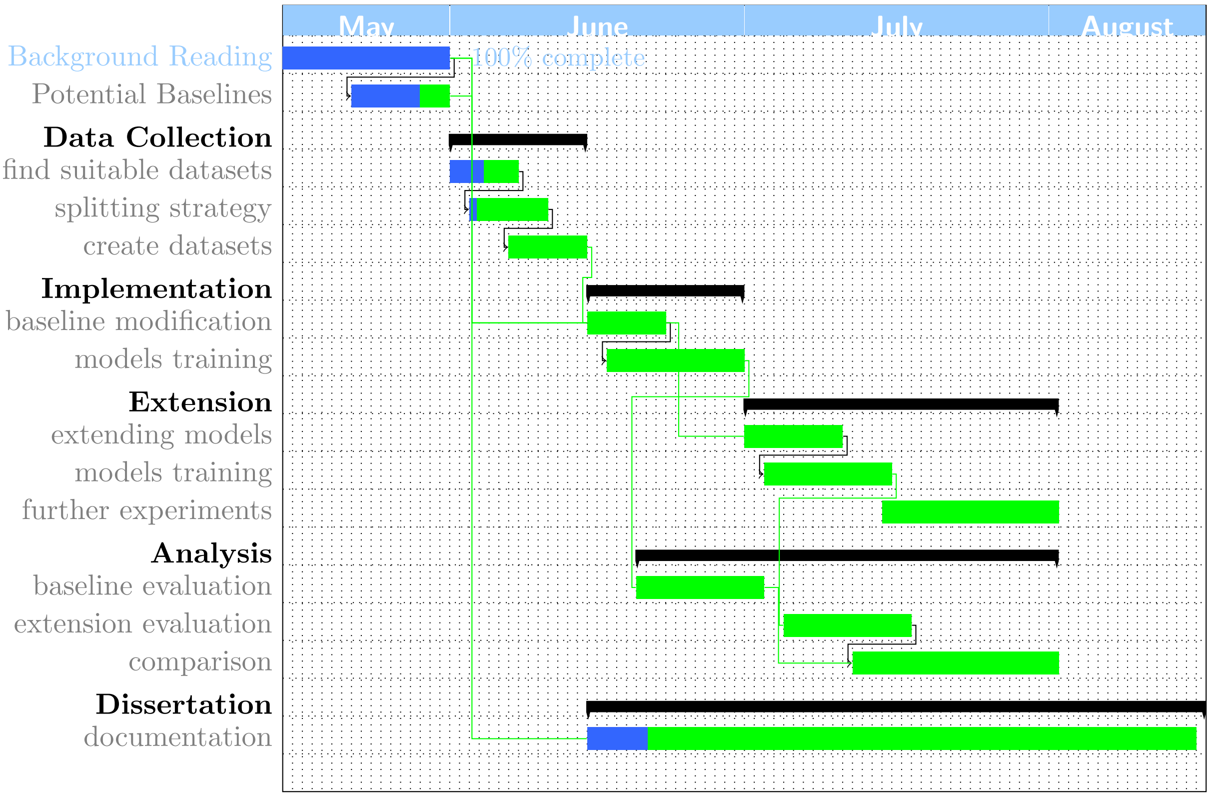

Figure 1: Gantt Chart of the activities defined for this project.

Milestone	Week	Description
M_1	1	Background Study completed
M_2	2	Data collection completed
M_3	4	Baseline model implementation completed
M_4	6	Extension for the time window completed
M_5	8	Evaluation completed
M_6	10	Submission of dissertation

Table 1: Milestones defined in this project.

Deliverable	Week	Description
D_1	2	useful processed dataset
D_2	5	Evaluation report on different baseline models
D_3	7	Evaluation report on multi-day window predictor
D_4	8	Comparison report on baseline models with the proposed model
D_5	10	Dissertation

Table 2: List of deliverables defined in this project.

References

[1] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: Recommendation models, techniques, and application fields. *Electronics*, 11(1), 2022.

[2] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Trans. Inf. Syst.*, 39(1), nov 2020.

[3] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks, 2016.

[4] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation, 2018.

[5] Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 565–573, New York, NY, USA, 2018. Association for Computing Machinery.

[6] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1441–1450, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[8] Zhiwei Liu, Yongjun Chen, Jia Li, Philip S. Yu, Julian McAuley, and Caiming Xiong. Contrastive self-supervised sequential recommendation with robust augmentation, 2021.

[9] Nikil Pancha, Andrew Zhai, Jure Leskovec, and Charles Rosenberg. Pinnerformer: Sequence modeling for user representation at pinterest. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3702–3712, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '19, page 2671–2679, New York, NY, USA, 2019. Association for Computing Machinery.

[11] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The youtube video recommendation system. In *Fourth ACM conference on Recommender systems*, 2010.

[12] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[13] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820, 2010.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[15] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Exploring data splitting strategies for the evaluation of recommendation models, 2020.