论文周报 | 推荐系统领域最新研究进展，含WWW, SIGIR, AAAI等顶会论文

本文精选了上周（0508-0514）最新发布的24篇推荐系统相关论文，主要研究方向包括大型语言模型赋能推荐系统、对话推荐系统、图推荐系统、隐私保护推荐系统、工业界推荐系统（来自谷歌、亚马逊、阿里）等。

以下整理了论文标题以及摘要，如感兴趣可移步原文精读。

1. Dual Intent Enhanced Graph Neural Network for Session-based New Item  Recommendation, WWW2023

2. NewsQuote: A Dataset Built on Quote Extraction and Attribution for  Expert Recommendation in Fact-Checking, AAAI2023

3. U-NEED: A Fine-grained Dataset for User Needs-Centric E-commerce  Conversational Recommendation, SIGIR2023

4. Graph Masked Autoencoder for Sequential Recommendation, SIGIR2023

5. Attacking Pre-trained Recommendation, SIGIR2023

6. Sim2Rec: A Simulator-based Decision-making Approach to Optimize  Real-World Long-term User Engagement in Sequential Recommender Systems, ICDE2023

7. FedPDD: A Privacy-preserving Double Distillation Framework for  Cross-silo Federated Recommendation, IJCNN2023

8. Multi-grained Hypergraph Interest Modeling for Conversational  Recommendation

9. Contrastive Enhanced Slide Filter Mixer for Sequential Recommendation

10. Multi-Task End-to-End Training Improves Conversational Recommendation, from Google

11. Learning to Personalize Recommendation based on Customers' Shopping  Intents, from Amazon

12. Privacy-Preserving Recommender Systems with Synthetic Query Generation  using Differentially Private Large Language Models, from Google

13. Explainable Recommender with Geometric Information Bottleneck

14. Recommender Systems with Generative Retrieval, from Google

15. Cross-domain Augmentation Networks for Click-Through Rate Prediction, from Alibaba

16. PerFedRec++: Enhancing Personalized Federated Recommendation with  Self-Supervised Pre-Training

17. How to Index Item IDs for Recommendation Foundation Models

18. Recommendation as Instruction Following: A Large Language Model  Empowered Recommendation Approach

19. A First Look at LLM-Powered Generative News Recommendation

20. Improving Implicit Feedback-Based Recommendation through Multi-Behavior  Alignment

21. DELTA: Direct Embedding Enhancement and Leverage Truncated Conscious  Attention for Recommendation System

22. WSFE: Wasserstein Sub-graph Feature Encoder for Effective User  Segmentation in Collaborative Filtering

23. Sparks of Artificial General Recommender (AGR): Early Experiments with  ChatGPT

24. Retraining A Graph-based Recommender with Interests Disentanglement

### **1. Dual Intent Enhanced Graph Neural Network for Session-based New Item  Recommendation, WWW2023**

Di Jin, Luzhi Wang, Yizhen Zheng, Guojie Song, Fei Jiang, Xiang Li, Wei Lin, Shirui Pan

https://arxiv.org/abs/2305.05848

Recommender systems are essential to various fields, e.g., e-commerce, e-learning, and streaming media. At present, graph neural networks (GNNs) for session-based recommendations normally can only recommend items existing in users' historical sessions. As a result, these GNNs have difficulty recommending items that users have never interacted with (new items), which leads to a phenomenon of information cocoon. Therefore, it is necessary to recommend new items to users. As there is no interaction between new items and users, we cannot include new items when building session graphs for GNN session-based recommender systems. Thus, it is challenging to recommend new items for users when using GNN-based methods. We regard this challenge as 'GNN Session-based New Item Recommendation (GSNIR)'. To solve this problem, we propose a dual-intent enhanced graph neural network for it. Due to the fact that new items are not tied to historical sessions, the users' intent is difficult to predict. We design a dual-intent network to learn user intent from an attention mechanism and the distribution of historical data respectively, which can simulate users' decision-making process in interacting with a new item. To solve the challenge that new items cannot be learned by GNNs, inspired by zero-shot learning (ZSL), we infer the new item representation in GNN space by using their attributes. By outputting new item probabilities, which contain recommendation scores of the corresponding items, the new items with higher scores are recommended to users. Experiments on two representative real-world datasets show the superiority of our proposed method. The case study from the real-world verifies interpretability benefits brought by the dual-intent module and the new item reasoning module. The code is available at Github: https://github.com/Ee1s/NirGNN

### **2. NewsQuote: A Dataset Built on Quote Extraction and Attribution for  Expert Recommendation in Fact-Checking, AAAI2023**

Wenjia Zhang, Lin Gui, Rob Procter, Yulan He

https://arxiv.org/abs/2305.04825

To enhance the ability to find credible evidence in news articles, we propose a novel task of expert recommendation, which aims to identify trustworthy experts on a specific news topic. To achieve the aim, we describe the construction of a novel NewsQuote dataset consisting of 24,031 quote-speaker pairs that appeared on a COVID-19 news corpus. We demonstrate an automatic pipeline for speaker and quote extraction via a BERT-based Question Answering model. Then, we formulate expert recommendations as document retrieval task by retrieving relevant quotes first as an intermediate step for expert identification, and expert retrieval by directly retrieving sources based on the probability of a query conditional on a candidate expert. Experimental results on NewsQuote show that document retrieval is more effective in identifying relevant experts for a given news topic compared to expert retrieval. https://github.com/WenjiaZh/NewsQuote

### **3. U-NEED: A Fine-grained Dataset for User Needs-Centric E-commerce  Conversational Recommendation, SIGIR2023**

Yuanxing Liu, Weinan Zhang, Baohua Dong, Yan Fan, Hang Wang, Fan Feng, Yifan Chen, Ziyu Zhuang, Hengbin Cui, Yongbin Li, Wanxiang Che

https://arxiv.org/abs/2305.04774

Conversational recommender systems (CRSs) aim to understand the information needs and preferences expressed in a dialogue to recommend suitable items to the user. Most of the existing conversational recommendation datasets are synthesized or simulated with crowdsourcing, which has a large gap with real-world scenarios. To bridge the gap, previous work contributes a dataset E-ConvRec, based on pre-sales dialogues between users and customer service staff in E-commerce scenarios. However, E-ConvRec only supplies coarse-grained annotations and general tasks for making recommendations in pre-sales dialogues. Different from that, we use real user needs as a clue to explore the E-commerce conversational recommendation in complex pre-sales dialogues, namely user needs-centric E-commerce conversational recommendation (UNECR).

In this paper, we construct a user needs-centric E-commerce conversational recommendation dataset (U-NEED) from real-world E-commerce scenarios. U-NEED consists of 3 types of resources: (i) 7,698 fine-grained annotated pre-sales dialogues in 5 top categories (ii) 333,879 user behaviors and (iii) 332,148 product knowledge tuples. To facilitate the research of UNECR, we propose 5 critical tasks: (i) pre-sales dialogue understanding (ii) user needs elicitation (iii) user needs-based recommendation (iv) pre-sales dialogue generation and (v) pre-sales dialogue evaluation. We establish baseline methods and evaluation metrics for each task. We report experimental results of 5 tasks on U-NEED. We also report results in 3 typical categories. Experimental results indicate that the challenges of UNECR in various categories are different.

### **4. Graph Masked Autoencoder for Sequential Recommendation, SIGIR2023**

Yaowen Ye, Lianghao Xia, Chao Huang

https://arxiv.org/abs/2305.04619

While some powerful neural network architectures (e.g., Transformer, Graph Neural Networks) have achieved improved performance in sequential recommendation with high-order item dependency modeling, they may suffer from poor representation capability in label scarcity scenarios. To address the issue of insufficient labels, Contrastive Learning (CL) has attracted much attention in recent methods to perform data augmentation through embedding contrasting for self-supervision. However, due to the hand-crafted property of their contrastive view generation strategies, existing CL-enhanced models i) can hardly yield consistent performance on diverse sequential recommendation tasks; ii) may not be immune to user behavior data noise. In light of this, we propose a simple yet effective Graph Masked AutoEncoder-enhanced sequential Recommender system (MAERec) that adaptively and dynamically distills global item transitional information for self-supervised augmentation. It naturally avoids the above issue of heavy reliance on constructing high-quality embedding contrastive views. Instead, an adaptive data reconstruction paradigm is designed to be integrated with the long-range item dependency modeling, for informative augmentation in sequential recommendation. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art baseline models and can learn more accurate representations against data noise and sparsity. Our implemented model code is available at https://github.com/HKUDS/MAERec

### **5. Attacking Pre-trained Recommendation, SIGIR2023**

Yiqing Wu, Ruobing Xie, Zhao Zhang, Yongchun Zhu, FuZhen Zhuang, Jie Zhou, Yongjun Xu, Qing He

https://arxiv.org/abs/2305.03995

Recently, a series of pioneer studies have shown the potency of pre-trained models in sequential recommendation, illuminating the path of building an omniscient unified pre-trained recommendation model for different downstream recommendation tasks. Despite these advancements, the vulnerabilities of classical recommender systems also exist in pre-trained recommendation in a new form, while the security of pre-trained recommendation model is still unexplored, which may threaten its widely practical applications. In this study, we propose a novel framework for backdoor attacking in pre-trained recommendation. We demonstrate the provider of the pre-trained model can easily insert a backdoor in pre-training, thereby increasing the exposure rates of target items to target user groups. Specifically, we design two novel and effective backdoor attacks: basic replacement and prompt-enhanced, under various recommendation pre-training usage scenarios. Experimental results on real-world datasets show that our proposed attack strategies significantly improve the exposure rates of target items to target users by hundreds of times in comparison to the clean model.

### **6. Sim2Rec: A Simulator-based Decision-making Approach to Optimize  Real-World Long-term User Engagement in Sequential Recommender Systems, ICDE2023**

Xiong-Hui Chen, Bowei He, Yang Yu, Qingyang Li, Zhiwei Qin, Wenjie Shang, Jieping Ye, Chen Ma

https://arxiv.org/abs/2305.04832

Long-term user engagement (LTE) optimization in sequential recommender systems (SRS) is shown to be suited by reinforcement learning (RL) which finds a policy to maximize long-term rewards. Meanwhile, RL has its shortcomings, particularly requiring a large number of online samples for exploration, which is risky in real-world applications. One of the appealing ways to avoid the risk is to build a simulator and learn the optimal recommendation policy in the simulator. In LTE optimization, the simulator is to simulate multiple users' daily feedback for given recommendations. However, building a user simulator with no reality-gap, i.e., can predict user's feedback exactly, is unrealistic because the users' reaction patterns are complex and historical logs for each user are limited, which might mislead the simulator-based recommendation policy. In this paper, we present a practical simulator-based recommender policy training approach, Simulation-to-Recommendation (Sim2Rec) to handle the reality-gap problem for LTE optimization. Specifically, Sim2Rec introduces a simulator set to generate various possibilities of user behavior patterns, then trains an environment-parameter extractor to recognize users' behavior patterns in the simulators. Finally, a context-aware policy is trained to make the optimal decisions on all of the variants of the users based on the inferred environment-parameters. The policy is transferable to unseen environments (e.g., the real world) directly as it has learned to recognize all various user behavior patterns and to make the correct decisions based on the inferred environment-parameters. Experiments are conducted in synthetic environments and a real-world large-scale ride-hailing platform, DidiChuxing. The results show that Sim2Rec achieves significant performance improvement, and produces robust recommendations in unseen environments.

### **7. FedPDD: A Privacy-preserving Double Distillation Framework for  Cross-silo Federated Recommendation, IJCNN2023**

Sheng Wan, Dashan Gao, Hanlin Gu, Daning Hu

https://arxiv.org/abs/2305.06272

Cross-platform recommendation aims to improve recommendation accuracy by gathering heterogeneous features from different platforms. However, such cross-silo collaborations between platforms are restricted by increasingly stringent privacy protection regulations, thus data cannot be aggregated for training. Federated learning (FL) is a practical solution to deal with the data silo problem in recommendation scenarios. Existing cross-silo FL methods transmit model information to collaboratively build a global model by leveraging the data of overlapped users. However, in reality, the number of overlapped users is often very small, thus largely limiting the performance of such approaches. Moreover, transmitting model information during training requires high communication costs and may cause serious privacy leakage. In this paper, we propose a novel privacy-preserving double distillation framework named FedPDD for cross-silo federated recommendation, which efficiently transfers knowledge when overlapped users are limited. Specifically, our double distillation strategy enables local models to learn not only explicit knowledge from the other party but also implicit knowledge from its past predictions. Moreover, to ensure privacy and high efficiency, we employ an offline training scheme to reduce communication needs and privacy leakage risk. In addition, we adopt differential privacy to further protect the transmitted information. The experiments on two real-world recommendation datasets, HetRec-MovieLens and Criteo, demonstrate the effectiveness of FedPDD compared to the state-of-the-art approaches.

### **8. Multi-grained Hypergraph Interest Modeling for Conversational  Recommendation**

Chenzhan Shang, Yupeng Hou, Wayne Xin Zhao, Yaliang Li, Jing Zhang

https://arxiv.org/abs/2305.04798

Conversational recommender system (CRS) interacts with users through multi-turn dialogues in natural language, which aims to provide high-quality recommendations for user's instant information need. Although great efforts have been made to develop effective CRS, most of them still focus on the contextual information from the current dialogue, usually suffering from the data scarcity issue. Therefore, we consider leveraging historical dialogue data to enrich the limited contexts of the current dialogue session.

In this paper, we propose a novel multi-grained hypergraph interest modeling approach to capture user interest beneath intricate historical data from different perspectives. As the core idea, we employ hypergraph to represent complicated semantic relations underlying historical dialogues. In our approach, we first employ the hypergraph structure to model users' historical dialogue sessions and form a session-based hypergraph, which captures coarse-grained, session-level relations. Second, to alleviate the issue of data scarcity, we use an external knowledge graph and construct a knowledge-based hypergraph considering fine-grained, entity-level semantics. We further conduct multi-grained hypergraph convolution on the two kinds of hypergraphs, and utilize the enhanced representations to develop interest-aware CRS. Extensive experiments on two benchmarks ReDial and TG-ReDial validate the effectiveness of our approach on both recommendation and conversation tasks. Code is available at: https://github.com/RUCAIBox/MHIM.

### **9. Contrastive Enhanced Slide Filter Mixer for Sequential Recommendation**

Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Junhua Fang, Guanfeng Liu, Yanchi Liu, Victor S. Sheng, Xiaofang Zhou

https://arxiv.org/abs/2305.04322

Sequential recommendation (SR) aims to model user preferences by capturing behavior patterns from their item historical interaction data. Most existing methods model user preference in the time domain, omitting the fact that users' behaviors are also influenced by various frequency patterns that are difficult to separate in the entangled chronological items. However, few attempts have been made to train SR in the frequency domain, and it is still unclear how to use the frequency components to learn an appropriate representation for the user. To solve this problem, we shift the viewpoint to the frequency domain and propose a novel Contrastive Enhanced SLIde Filter MixEr for Sequential Recommendation, named SLIME4Rec. Specifically, we design a frequency ramp structure to allow the learnable filter slide on the frequency spectrums across different layers to capture different frequency patterns. Moreover, a Dynamic Frequency Selection (DFS) and a Static Frequency Split (SFS) module are proposed to replace the self-attention module for effectively extracting frequency information in two ways. DFS is used to select helpful frequency components dynamically, and SFS is combined with the dynamic frequency selection module to provide a more fine-grained frequency division. Finally, contrastive learning is utilized to improve the quality of user embedding learned from the frequency domain. Extensive experiments conducted on five widely used benchmark datasets demonstrate our proposed model performs significantly better than the state-of-the-art approaches. Our code is available at https://github.com/sudaada/SLIME4Rec

### **10. Multi-Task End-to-End Training Improves Conversational Recommendation, from Google**

Naveen Ram, Dima Kuzmin, Ellie Ka In Chio, Moustafa Farid Alzantot, Santiago Ontanon, Ambarish Jash, Judith Yue Li

https://arxiv.org/abs/2305.06218

In this paper, we analyze the performance of a multitask end-to-end transformer model on the task of conversational recommendations, which aim to provide recommendations based on a user's explicit preferences expressed in dialogue. While previous works in this area adopt complex multi-component approaches where the dialogue management and entity recommendation tasks are handled by separate components, we show that a unified transformer model, based on the T5 text-to-text transformer model, can perform competitively in both recommending relevant items and generating conversation dialogue. We fine-tune our model on the ReDIAL conversational movie recommendation dataset, and create additional training tasks derived from MovieLens (such as the prediction of movie attributes and related movies based on an input movie), in a multitask learning setting. Using a series of probe studies, we demonstrate that the learned knowledge in the additional tasks is transferred to the conversational setting, where each task leads to a 9%-52% increase in its related probe score.

### **11. Learning to Personalize Recommendation based on Customers' Shopping  Intents, from Amazon**

Xin Shen, Jiaying Shi, Sungro Yoon, Jon Katzur, Hanbo Wang, Jim Chan, Jin Li

https://arxiv.org/abs/2305.05279

Understanding the customers' high level shopping intent, such as their desire to go camping or hold a birthday party, is critically important for an E-commerce platform; it can help boost the quality of shopping experience by enabling provision of more relevant, explainable, and diversified recommendations. However, such high level shopping intent has been overlooked in the industry due to practical challenges. In this work, we introduce Amazon's new system that explicitly identifies and utilizes each customer's high level shopping intents for personalizing recommendations. We develop a novel technique that automatically identifies various high level goals being pursued by the Amazon customers, such as "go camping", and "preparing for a beach party". Our solution is in a scalable fashion (in 14 languages across 21 countries). Then a deep learning model maps each customer's online behavior, e.g. product search and individual item engagements, into a subset of high level shopping intents. Finally, a realtime ranker considers both the identified intents as well as the granular engagements to present personalized intent-aware recommendations. Extensive offline analysis ensures accuracy and relevance of the new recommendations and we further observe an 10% improvement in the business metrics. This system is currently serving online traffic at http://amazon.com/, powering several production features, driving significant business impacts.

### **12. Privacy-Preserving Recommender Systems with Synthetic Query Generation  using Differentially Private Large Language Models, from Google**

Aldo Gael Carranza, Rezsa Farahani, Natalia Ponomareva, Alex Kurakin, Matthew Jagielski, Milad Nasr

https://arxiv.org/abs/2305.05973

We propose a novel approach for developing privacy-preserving large-scale recommender systems using differentially private (DP) large language models (LLMs) which overcomes certain challenges and limitations in DP training these complex systems. Our method is particularly well suited for the emerging area of LLM-based recommender systems, but can be readily employed for any recommender systems that process representations of natural language inputs. Our approach involves using DP training methods to fine-tune a publicly pre-trained LLM on a query generation task. The resulting model can generate private synthetic queries representative of the original queries which can be freely shared for any downstream non-private recommendation training procedures without incurring any additional privacy cost. We evaluate our method on its ability to securely train effective deep retrieval models, and we observe significant improvements in their retrieval quality without compromising query-level privacy guarantees compared to methods where the retrieval models are directly DP trained.

### **13. Explainable Recommender with Geometric Information Bottleneck**

Hanqi Yan, Lin Gui, Menghan Wang, Kun Zhang, Yulan He

https://arxiv.org/abs/2305.05331

Explainable recommender systems can explain their recommendation decisions, enhancing user trust in the systems. Most explainable recommender systems either rely on human-annotated rationales to train models for explanation generation or leverage the attention mechanism to extract important text spans from reviews as explanations. The extracted rationales are often confined to an individual review and may fail to identify the implicit features beyond the review text. To avoid the expensive human annotation process and to generate explanations beyond individual reviews, we propose to incorporate a geometric prior learnt from user-item interactions into a variational network which infers latent factors from user-item reviews. The latent factors from an individual user-item pair can be used for both recommendation and explanation generation, which naturally inherit the global characteristics encoded in the prior knowledge. Experimental results on three e-commerce datasets show that our model significantly improves the interpretability of a variational recommender using the Wasserstein distance while achieving performance comparable to existing content-based recommender systems in terms of recommendation behaviours.

### **14. Recommender Systems with Generative Retrieval, from Google**

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, Maheswaran Sathiamoorthy

https://arxiv.org/abs/2305.05065

Modern recommender systems leverage large-scale retrieval models consisting of two stages: training a dual-encoder model to embed queries and candidates in the same space, followed by an Approximate Nearest Neighbor (ANN) search to select top candidates given a query's embedding. In this paper, we propose a new single-stage paradigm: a generative retrieval model which autoregressively decodes the identifiers for the target candidates in one phase. To do this, instead of assigning randomly generated atomic IDs to each item, we generate Semantic IDs: a semantically meaningful tuple of codewords for each item that serves as its unique identifier. We use a hierarchical method called RQ-VAE to generate these codewords. Once we have the Semantic IDs for all the items, a Transformer based sequence-to-sequence model is trained to predict the Semantic ID of the next item. Since this model predicts the tuple of codewords identifying the next item directly in an autoregressive manner, it can be considered a generative retrieval model. We show that our recommender system trained in this new paradigm improves the results achieved by current SOTA models on the Amazon dataset. Moreover, we demonstrate that the sequence-to-sequence model coupled with hierarchical Semantic IDs offers better generalization and hence improves retrieval of cold-start items for recommendations.

### **15. Cross-domain Augmentation Networks for Click-Through Rate Prediction, from Alibaba**

Xu Chen, Zida Cheng, Shuai Xiao, Xiaoyi Zeng, Weilin Huang

https://arxiv.org/abs/2305.03953

Data sparsity is an important issue for click-through rate (CTR) prediction, particularly when user-item interactions is too sparse to learn a reliable model. Recently, many works on cross-domain CTR (CDCTR) prediction have been developed in an effort to leverage meaningful data from a related domain. However, most existing CDCTR works have an impractical limitation that requires homogeneous inputs (i.e. shared feature fields) across domains, and CDCTR with heterogeneous inputs (i.e. varying feature fields) across domains has not been widely explored but is an urgent and important research problem. In this work, we propose a cross-domain augmentation network (CDAnet) being able to perform knowledge transfer between two domains with heterogeneous inputs. Specifically, CDAnet contains a designed translation network and an augmentation network which are trained sequentially. The translation network is able to compute features from two domains with heterogeneous inputs separately by designing two independent branches, and then learn meaningful cross-domain knowledge using a designed cross-supervised feature translator. Later the augmentation network encodes the learned cross-domain knowledge via feature translation performed in the latent space and fine-tune the model for final CTR prediction. Through extensive experiments on two public benchmarks and one industrial production dataset, we show CDAnet can learn meaningful translated features and largely improve the performance of CTR prediction. CDAnet has been conducted online A/B test in image2product retrieval at Taobao app over 20days, bringing an absolute 0.11 point CTR improvement and a relative 1.26% GMV increase.

### **16. PerFedRec++: Enhancing Personalized Federated Recommendation with  Self-Supervised Pre-Training**

Sichun Luo, Yuanzhang Xiao, Xinyi Zhang, Yang Liu, Wenbo Ding, Linqi Song

https://arxiv.org/abs/2305.06622

Federated recommendation systems employ federated learning techniques to safeguard user privacy by transmitting model parameters instead of raw user data between user devices and the central server. Nevertheless, the current federated recommender system faces challenges such as heterogeneity and personalization, model performance degradation, and communication bottleneck. Previous studies have attempted to address these issues, but none have been able to solve them simultaneously.

In this paper, we propose a novel framework, named PerFedRec++, to enhance the personalized federated recommendation with self-supervised pre-training. Specifically, we utilize the privacy-preserving mechanism of federated recommender systems to generate two augmented graph views, which are used as contrastive tasks in self-supervised graph learning to pre-train the model. Pre-training enhances the performance of federated models by improving the uniformity of representation learning. Also, by providing a better initial state for federated training, pre-training makes the overall training converge faster, thus alleviating the heavy communication burden. We then construct a collaborative graph to learn the client representation through a federated graph neural network. Based on these learned representations, we cluster users into different user groups and learn personalized models for each cluster. Each user learns a personalized model by combining the global federated model, the cluster-level federated model, and its own fine-tuned local model. Experiments on three real-world datasets show that our proposed method achieves superior performance over existing methods.

### **17. How to Index Item IDs for Recommendation Foundation Models**

Wenyue Hua, Shuyuan Xu, Yingqiang Ge, Yongfeng Zhang

https://arxiv.org/abs/2305.06569

Recommendation foundation model utilizes large language models (LLM) for recommendation by converting recommendation tasks into natural language tasks. It enables generative recommendation which directly generates the item(s) to recommend rather than calculating a ranking score for each and every candidate item in traditional recommendation models, simplifying the recommendation pipeline from multi-stage filtering to single-stage filtering. To avoid generating excessively long text when deciding which item(s) to recommend, creating LLM-compatible item IDs is essential for recommendation foundation models. In this study, we systematically examine the item indexing problem for recommendation foundation models, using P5 as the representative backbone model and replicating its results with various indexing methods. To emphasize the importance of item indexing, we first discuss the issues of several trivial item indexing methods, such as independent indexing, title indexing, and random indexing. We then propose four simple yet effective solutions, including sequential indexing, collaborative indexing, semantic (content-based) indexing, and hybrid indexing. Our reproducibility study of P5 highlights the significant influence of item indexing methods on the model performance, and our results on real-world datasets validate the effectiveness of our proposed solutions.

### **18. Recommendation as Instruction Following: A Large Language Model  Empowered Recommendation Approach**

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, Ji-Rong Wen

https://arxiv.org/abs/2305.07001

In the past decades, recommender systems have attracted much attention in both research and industry communities, and a large number of studies have been devoted to developing effective recommendation models. Basically speaking, these models mainly learn the underlying user preference from historical behavior data, and then estimate the user-item matching relationships for recommendations. Inspired by the recent progress on large language models (LLMs), we take a different approach to developing the recommendation models, considering recommendation as instruction following by LLMs. The key idea is that the preferences or needs of a user can be expressed in natural language descriptions (called instructions), so that LLMs can understand and further execute the instruction for fulfilling the recommendation task. Instead of using public APIs of LLMs, we instruction tune an open-source LLM (3B Flan-T5-XL), in order to better adapt LLMs to recommender systems. For this purpose, we first design a general instruction format for describing the preference, intention, task form and context of a user in natural language. Then we manually design 39 instruction templates and automatically generate a large amount of user-personalized instruction data (252K instructions) with varying types of preferences and intentions. To demonstrate the effectiveness of our approach, we instantiate the instruction templates into several widely-studied recommendation (or search) tasks, and conduct extensive experiments on these tasks with real-world datasets. Experiment results show that the proposed approach can outperform several competitive baselines, including the powerful GPT-3.5, on these evaluation tasks. Our approach sheds light on developing more user-friendly recommender systems, in which users can freely communicate with the system and obtain more accurate recommendations via natural language instructions.

### **19. A First Look at LLM-Powered Generative News Recommendation**

Qijiong Liu, Nuo Chen, Tetsuya Sakai, Xiao-Ming Wu

https://arxiv.org/abs/2305.06566

Personalized news recommendation systems have become essential tools for users to navigate the vast amount of online news content, yet existing news recommenders face significant challenges such as the cold-start problem, user profile modeling, and news content understanding. Previous works have typically followed an inflexible routine to address a particular challenge through model design, but are limited in their ability to understand news content and capture user interests. In this paper, we introduce GENRE, an LLM-powered generative news recommendation framework, which leverages pretrained semantic knowledge from large language models to enrich news data. Our aim is to provide a flexible and unified solution for news recommendation by moving from model design to prompt design. We showcase the use of GENRE for personalized news generation, user profiling, and news summarization. Extensive experiments with various popular recommendation models demonstrate the effectiveness of GENRE. We will publish our code and data for other researchers to reproduce our work.

### **20. Improving Implicit Feedback-Based Recommendation through Multi-Behavior  Alignment**

Xin Xin, Xiangyuan Liu, Hanbing Wang, Pengjie Ren, Zhumin Chen, Jiahuan Lei, Xinlei Shi, Hengliang Luo, Joemon Jose, Maarten de Rijke, Zhaochun Ren

https://arxiv.org/abs/2305.05585

Recommender systems that learn from implicit feedback often use large volumes of a single type of implicit user feedback, such as clicks, to enhance the prediction of sparse target behavior such as purchases. Using multiple types of implicit user feedback for such target behavior prediction purposes is still an open question. Existing studies that attempted to learn from multiple types of user behavior often fail to: (i) learn universal and accurate user preferences from different behavioral data distributions, and (ii) overcome the noise and bias in observed implicit user feedback. To address the above problems, we propose multi-behavior alignment (MBA), a novel recommendation framework that learns from implicit feedback by using multiple types of behavioral data. We conjecture that multiple types of behavior from the same user (e.g., clicks and purchases) should reflect similar preferences of that user. To this end, we regard the underlying universal user preferences as a latent variable. The variable is inferred by maximizing the likelihood of multiple observed behavioral data distributions and, at the same time, minimizing the Kullback-Leibler divergence (KL-divergence) between user models learned from auxiliary behavior (such as clicks or views) and the target behavior separately. MBA infers universal user preferences from multi-behavior data and performs data denoising to enable effective knowledge transfer. We conduct experiments on three datasets, including a dataset collected from an operational e-commerce platform. Empirical results demonstrate the effectiveness of our proposed method in utilizing multiple types of behavioral data to enhance the prediction of the target behavior.

### **21. DELTA: Direct Embedding Enhancement and Leverage Truncated Conscious  Attention for Recommendation System**

Chen Zhu, Liang Du, Xin Wang, Wenwu Zhu

https://arxiv.org/abs/2305.04891

Click-Through Rate (CTR) prediction is the most critical task in product and content recommendation, and learning effective feature interaction is the key challenge to exploiting user preferences for products. Some recent research works focus on investigating more sophisticated feature interactions based on soft attention or gate mechanism, while some redundant or contradictory feature combinations are still introduced. According to Global Workspace Theory in conscious processing, human clicks on advertisements ''consciously'': only a specific subset of product features are considered, and the rest are not involved in conscious processing. Therefore, we propose a CTR model that Directly Enhances the embeddings and Leverages Truncated Conscious Attention during feature interaction, termed DELTA, which contains two key components: (I) conscious truncation module (CTM), which utilizes curriculum learning to apply adaptive truncation on attention weights to select the most critical feature combinations; (II) direct embedding enhancement module (DEM), which directly and independently propagates gradient from the loss layer to the embedding layer to enhance the crucial embeddings via linear feature crossing without introducing any extra cost during inference. Extensive experiments on five challenging CTR datasets demonstrate that DELTA achieves cutting-edge performance among current state-of-the-art CTR methods.

### **22. WSFE: Wasserstein Sub-graph Feature Encoder for Effective User  Segmentation in Collaborative Filtering**

Yankai Chen, Yifei Zhang, Menglin Yang, Zixing Song, Chen Ma, Irwin King

https://arxiv.org/abs/2305.04410

Maximizing the user-item engagement based on vectorized embeddings is a standard procedure of recent recommender models. Despite the superior performance for item recommendations, these methods however implicitly deprioritize the modeling of user-wise similarity in the embedding space; consequently, identifying similar users is underperforming, and additional processing schemes are usually required otherwise. To avoid thorough model re-training, we propose WSFE, a model-agnostic and training-free representation encoder, to be flexibly employed on the fly for effective user segmentation. Underpinned by the optimal transport theory, the encoded representations from WSFE present a matched user-wise similarity/distance measurement between the realistic and embedding space. We incorporate WSFE into six state-of-the-art recommender models and conduct extensive experiments on six real-world datasets. The empirical analyses well demonstrate the superiority and generality of WSFE to fuel multiple downstream tasks with diverse underlying targets in recommendation.

### **23. Sparks of Artificial General Recommender (AGR): Early Experiments with  ChatGPT**

Guo Lin, Yongfeng Zhang

https://arxiv.org/abs/2305.04518

This study investigates the feasibility of developing an Artificial General Recommender (AGR), facilitated by recent advancements in Large Language Models (LLMs). An AGR comprises both conversationality and universality to engage in natural dialogues and generate recommendations across various domains. We propose ten fundamental principles that an AGR should adhere to, each with its corresponding testing protocols. We proceed to assess whether ChatGPT, a sophisticated LLM, can comply with the proposed principles by engaging in recommendation-oriented dialogues with the model while observing its behavior. Our findings demonstrate the potential for ChatGPT to serve as an AGR, though several limitations and areas for improvement are identified.

### **24. Retraining A Graph-based Recommender with Interests Disentanglement**

Yitong Ji, Aixin Sun, Jie Zhang

https://arxiv.org/abs/2305.03624

In a practical recommender system, new interactions are continuously observed. Some interactions are expected, because they largely follow users' long-term preferences. Some other interactions are indications of recent trends in user preference changes or marketing positions of new items. Accordingly, the recommender needs to be periodically retrained or updated to capture the new trends, and yet not to forget the long-term preferences. In this paper, we propose a novel and generic retraining framework called Disentangled Incremental Learning (DIL) for graph-based recommenders. We assume that long-term preferences are well captured in the existing model, in the form of model parameters learned from past interactions. New preferences can be learned from the user-item bipartite graph constructed using the newly observed interactions. In DIL, we design an Information Extraction Module to extract historical preferences from the existing model. Then we blend the historical and new preferences in the form of node embeddings in the new graph, through a Disentanglement Module. The essence of the disentanglement module is to decorrelate the historical and new preferences so that both can be well captured, via carefully designed losses. Through experiments on three benchmark datasets, we show the effectiveness of DIL in capturing dynamics of useritem interactions. We also demonstrate the robustness of DIL by attaching it to two base models - LightGCN and NGCF.