

LongRetriever: Towards Ultra-Long Sequence based Candidate Retrieval for Recommendation

Qin Ren
renqin.97@bytedance.com
ByteDance
Beijing, China

Zheng Chai[†]
chaizheng.cz@bytedance.com
ByteDance
Hangzhou, China

Xijun Xiao
xiaoxijun@bytedance.com
ByteDance
Beijing, China

Yuchao Zheng[†]
zhengyuchao.yc@bytedance.com
ByteDance
Beijing, China

Di Wu
di.wu@bytedance.com
ByteDance
Beijing, China

Abstract

Precisely modeling user ultra-long sequences is critical for industrial recommender systems. Current approaches predominantly focus on leveraging ultra-long sequences in the ranking stage, whereas research for the candidate retrieval stage remains under-explored. This paper presents LongRetriever, a practical framework for incorporating ultra-long sequences into the retrieval stage of recommenders. Specifically, we propose in-context training and multi-context retrieval, which enable candidate-specific interaction between user sequence and candidate item, and ensure training-serving consistency under the search-based paradigm. Extensive online A/B testing conducted on a large-scale e-commerce platform demonstrates statistically significant improvements, confirming the framework's effectiveness. Currently, LongRetriever has been fully deployed in the platform, impacting billions of users.

CCS Concepts

• Information systems → Recommender systems.

Keywords

Long Sequence Modeling, Candidate Matching, Industrial Recommenders

ACM Reference Format:

Qin Ren, Zheng Chai[†], Xijun Xiao, Yuchao Zheng[†], and Di Wu. 2018. LongRetriever: Towards Ultra-Long Sequence based Candidate Retrieval for Recommendation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXXX.XXXXXXX>

[†]Corresponding Authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Retrieving highly relevant items from a massive item repository within constrained timeframes presents a significant challenge in industrial recommendation systems [3, 13]. The predominant solution employs a dual-stage methodology [6, 7], wherein an initial *candidate retrieval* stage is dedicated to narrowing down the candidate set of items, enabling subsequent deployment of more precise yet computationally intensive models during the *ranking* stage.

Generally, popularity-based filtering and collaborative filtering have been widely adopted in candidate retrieval stage. With the advancement of deep learning, embedding-based retrieval (EBR) has emerged as the predominant paradigm [8, 10]. EBR utilizes a dual-tower deep neural architecture to encode user and item features into dense vector representations, respectively. It optimizes a distance metric (e.g., cosine similarity) between these vectors within a discriminative learning framework, distinguishing relevant items (positives) from irrelevant ones (negatives). During deployment, item vectors are precomputed and indexed using approximate nearest neighbor (ANN) search techniques [1] (e.g., HNSW [11]), enabling efficient top-k item retrieval in sub-linear time for online serving.

Despite a favorable trade-off between efficiency and efficacy, constrained by computational resources and online latency requirements, EBR typically processes only short user sequences (e.g., the most recent 200 behaviors) in the User Encoder. Besides, due to the separated dual-tower learning paradigm, it is generally difficult for the candidate item to interact with user-side historical behavioral sequences, leveraging a remarkable margin for understanding user interest deeply. Precisely modeling user sequences is essential to current recommenders, in which typical approaches generally adopt a target-attention mechanism for modeling user interest, and the most representative approaches include DIN [15], DIEN [14], CAN [2], etc. With the advancement of computational capabilities and the ever-growing accumulation of user online behaviors, recent research has shifted toward achieving more comprehensive modeling of ultra-long sequences in recommender systems [5, 12].

Despite the popularity, it is noted that current long-sequence modeling is mostly developed for the ranking stage in recommenders. How to leverage the power of ultra-long sequence for enhancing EBR remains a notable gap in industrial recommenders. Generally, there are the following challenges in leveraging ultra-long sequence in candidate retrieval:

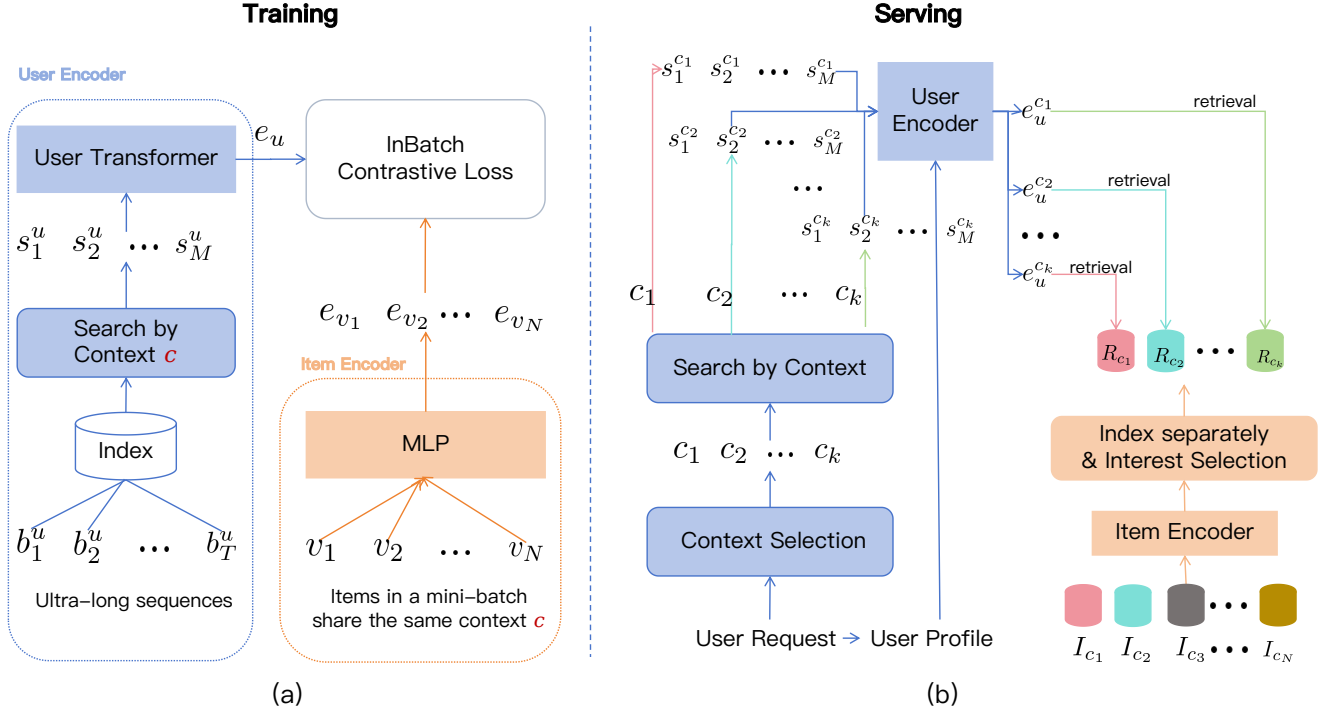


Figure 1: LongRetriever Model Framework.

- Adequate interaction between the ultra-long sequence and candidate item. Note that the popular target attention-based approaches are applicable in ranking stage due to the limited number of candidate items (around hundreds in industry). While in the matching stage, the candidate items scale to the order of hundreds of billions, making the interaction between the ultra-long sequence and candidate item intractable in practice.
- Diverse while precise interest modeling with controllable learning capability with ultra-long sequences. Traditional dual-tower paradigms generally encode user representation as a single vector which bottlenecks modeling diverse interests at industrial scale, and existing multi-interest modules [4, 9] generally lack concrete semantics, leading to reduced controllability.

To tackle the above challenges, in this paper, we present *LongRetriever*, a controllable and accurate diverse-interests learning framework for candidate matching with ultra-long sequences. LongRetriever defines mutually exclusive item categories as interpretable user interests and conducts independent candidate training for each interest under shared model parameters. It employs a search-based mechanism to filter relevant subsequences from the user’s lifelong behavior sequence for each interest. This approach represents users as multiple dense vectors across different interest domains without increasing training overhead. The utilization of the lifelong user behavior sequence, which demonstrates strong effectiveness in the ranking stage, also enables LongRetriever to achieve more accurate expression within each interest in the candidate matching stage. Overall, the contributions are mainly summarized as follows:

- We devised LongRetriever, a controllable multi-interest learning framework based on long sequence. To the best of our knowledge, it makes the first attempt to introduce ultra-long sequence to the candidate matching stage.
- We designed novel in-context training and multi-context serving for introducing user-candidate interaction and ensuring training-serving consistency under the search-based paradigm. Combined with interest selection, this renders the entire retrieval fully interpretable and controllable.
- Extensive online A/B testing on a super e-commerce platform with billions of users validates LongRetriever’s effectiveness. The framework is now fully deployed at industrial scale, impacting billions of users.

2 METHODOLOGY

2.1 Background: Embedding-based Retrieval

Let U and I denote the user and item sets respectively. Given a user $u \in U$ with raw behavior sequence $B_u = [b_1^u, b_2^u, \dots, b_T^u]$, user basic features P_u including user profiles and context features, and a target item $v \in I$. Embedding-based retrieval frameworks employ a user encoder f and an item encoder g to project user and item features into low-dimensional dense vectors: $e_u = f(B_u, P_u)$ and $e_v = g(v)$. The relevance of user u to item v , denoted by r_{uv} , is quantified using a distance metric (typically cosine similarity) between x_u and x_v . During training, the relevance scores r_{uv} are optimized via an in-batch contrastive loss to distinguish between items interacted by u and non-interacted items within the same

mini-batch. During deployment, all item vectors $\{e_v\} v \in I$ are precomputed and indexed using approximate nearest neighbor search techniques. Consequently, only the user vector e_u requires real-time computation, and the ANN system can retrieve the k nearest items to e_u from the item repository in sub-linear time.

2.2 Overall Framework

Figure 1 illustrates the overall architecture of our LongRetriever model. Generally, given the list of user lifelong behaviors $B_u = [b_1^u, b_2^u, \dots, b_T^u]$ where T is the length of user behaviors. LongRetriever first employs category matching to filter the most recent L behaviors from B_u that exactly match the candidate item's context, similar to the Hard-Search method in SIM[12]. Here, the context can be any criteria that can be relied on in practice. In this paper, we adopt item's category for hard search. After search based on context, the new user subsequence is denoted as $S_u = [s_1^u, s_2^u, \dots, s_M^u]$.

LongRetriever subsequently employs a user transformer model as the user encoder f to extract the user vector e_u . The input tokens to this Transformer comprise: - A $[CLS]$ token to aggregate information across all tokens. - K tokens derived from user profile features $P_u = [p_1^u, p_2^u, \dots, p_K^u]$ to enhance fine-grained interactions between the behavior sequence and basic user features. - L filtered sub-behaviors S_u . After adding learnable positional encodings and token type embeddings to the input tokens, LongRetriever utilizes a standard Pre-Layer Normalization Transformer to model the tokens. The final representation of the $[CLS]$ token is adopted as the user representation vector e_u :

$$e_u = \text{Transformer}([CLS, p_1^u, p_2^u, \dots, p_K^u, s_1^u, s_2^u, \dots, s_M^u]) \quad (1)$$

It is noteworthy that EBR's training requires independent computation of user and item encoders, to meet the requirements of the ANN paradigm during deployment. Consequently, candidate item features are excluded from our User Transformer.

Based on the model structure design, there are two key components designed in LongRetriever, In-Context Training and Multi-Context Retrieval.

2.3 In-Context Training

Directly leveraging the search-based long sequence yields the following loss function:

$$\mathcal{L}(e_u) = -\log \left(\frac{e^{e_u^T e_v}}{e^{e_u^T e_v} + \sum_{v' \in \mathcal{B}} e^{e_u^T e_{v'}}} \right) \quad (2)$$

where the \mathcal{B} denotes the in-batch negative item set, the v denotes the positive item and the v' indicates an arbitrary negative item.

While leveraging the lifelong behavior sequence resulted in significant improvements in offline evaluation metrics, this improvement is proved unreliable, because S_u in e_u implicitly incorporates partial information about the candidate item. This introduces data leakage under the in-batch contrastive objective, where instances influence each other. Consider an instance of interaction between user u and item v , where the context of S_u and v , i.e., the category of the items, is "Food." This (u, v) pair is treated as a positive example, whereas pairs of u with items from other instances in the current mini-batch (which may belong to categories like "Clothes"

or "Pets") are treated as negative examples. Under these conditions, the model can readily reduce the in-batch contrastive loss by leveraging whether the category of S_u matches that of v , thereby artificially inflating offline metrics.

To mitigate this data leakage, LongRetriever adopts an In-Context Training strategy. In-Context Training fundamentally involves restructuring samples within each mini-batch, using caching mechanisms to ensure that the subsequence S_u and the candidate item v within the same mini-batch share only one interest category.

Before In-Context Training, EBR's objective equates to discriminating whether a user would purchase items in the "Food" category over the entire sample space, given the category-specific sequence S_u . After In-Context Training, EBR's learning objective changes to discriminating whether a user would purchase a candidate item over the entire "Food" category sample space, given the category-specific sequence S_u .

2.4 Multi-Context Retrieval

The In-Context Training enables user-candidate interaction during training. While in serving stage, due to the separation inference of the user and item tower, it is of significance to infer several categories that users may be interested in to enable multi-context retrieval based on user's ultra-long behavior sequence. Here, two necessary stages are devised for user interest estimation and multi-interest retrieval.

Interest Selection and Representations. LongRetriever selects specific interests based on user historical behaviors and executes the user encoder multiple times to generate distinct user representations for each interest. Due to computational constraints, the number of interests selected per request should not be excessive. To this, we provide a simple automated interest selection strategy, *Random in Top*. It computes a user's engagement score for each interest c_i by aggregating historical behavioral data across interests using time-weighted summation:

$$\text{Score}_{c_i} = \sum_{b_i^u \in B_u} \frac{\text{sign}(c_i = c_{b_i^u})}{t_{b_i^u}} \quad (3)$$

where $c_{b_i^u}$ denotes the interest category associated with behavior b_i , and $t_{b_i^u}$ represents the time difference (in days) between the occurrence timestamp of behavior b_i and the current request timestamp. *Random in Top* first selects the $\text{top} - M$ interests by engagement score from the entire interest set, then randomly selects N interests from this subset for each user request.

Multi-Interest Retrieval. Following In-Context Training, the user vector is trained exclusively to mine relevant items within a specific interest domain. Consequently, LongRetriever partitions the entire item repository into multiple independent sub-repositories based on category and constructs indexes for each sub-repository. Upon receiving a user request, LongRetriever executes multiple retrievals across sub-repositories associated with selected interests and merges the results using predefined strategies.

Table 1: Online A/B Test Results on Business Metrics

PV	UV CTR	UV CVR	Orders Per User	Exposed Categories Per User	Clicked Categories Per User
+0.62%	+0.17%	+1.33%	+1.70%	+0.14%	+1.39%

3 EXPERIMENT

3.1 Experimental Setting

We evaluate LongRetriever in our shopping mall, a real-world large-scale industrial recommendation scenario. Unlike the targeted ranking stage, industrial matching systems are highly complex, typically comprising dozens of distinct retrieval strategies. Consequently, offline metrics alone are insufficient for evaluating retrieval effectiveness. We first train models on industrial datasets, then conduct a 7-day online A/B test with real users for evaluation. The training set comprises user interaction logs from our shopping mall, covering 900 million users, 150 million items, and over 10 billion samples spanning 400 consecutive days. Each sample includes user demographic features (e.g., user ID, gender), candidate item, a 50-behavior subsequence filtered from the user’s raw lifelong behavior sequence—which averages over 20,000 interactions—based on the candidate item’s category, and a label indicating whether there has been an interaction or not. We benchmark LongRetriever against MIND[9] — a widely adopted multi-interest retrieval strategy deployed in our production environment. LongRetriever adopts the Random 5 in Top 20 strategy to obtain 5 interests.

Table 2: Online A/B Test Results On Intermediate Metrics

Model	AER	UER	CTR*CVR
MIND	31.82%	4.58%	6.89%
LongRetriever	18.72%	9.29%	7.92%

Table 3: Ablation Study on Random Interests

Model	AER	UER	CTR*CVR
Random5 in Top20	18.72%	9.29%	7.92%
Top5	30.43%	7.23%	8.29%

Table 4: Ablation Study on Lifelong Behavior Sequence

Model	AER	UER	CTR*CVR
LongRetriever	18.72%	9.29%	7.92%
w/o Long Sequence	11.45%	5.92%	5.33%

3.2 Experimental results

Table 1 summarizes the performance of LongRetriever compared to MIND in our shopping mall. Due to commercial policies, we report only relative increases for business metrics and bold items

with statistically significant results ($p < 0.05$). In terms of convert efficiency, LongRetriever improved item page views (PV) by 0.62%, unique visitor (UV) click-through rate (CTR) by 0.17%, and UV conversion rate (CVR) by 1.33% compared to MIND. This collectively resulted in a 1.7% increase in average orders per user. Regarding ecosystem development, LongRetriever enhanced recommendation diversity by increasing average exposed categories per user by 0.14% and average clicked categories per user by 1.39%.

Table 2 presents intermediate metrics of retrieved items, revealing potential reasons for the business metric improvements in Table 1. All exposure ratio (AER) is used to measure the preference degree of a recommender for all results retrieved by the model. Unique exposure ratio (UER), on the other hand, is used to measure the preference degree of the recommendation system for items that can only be retrieved by this model. LongRetriever exhibits a significantly lower AER than MIND, primarily because MIND retrieves from the full item repository per user vector, whereas LongRetriever uses only 5 categories for Multi-Context Retrieval, severely limiting its candidate scope. However, LongRetriever achieves a 4.71% higher UER compared to MIND. This indicates that LongRetriever retrieves items highly distinct from the items retrieved by other existing strategies. Furthermore, the unique conversion efficiency of LongRetriever, measured by $CTR * CVR$, improves by 1.03% over MIND, thus contributing to business metric gains.

3.3 Ablation study

Random Interests. Table 3 compares two different interest selection strategies. User interest distributions typically exhibit long-tailed characteristics. When retaining only the top 5 interests in interest selection, the AER increased by 11.71% compared to the Random in Top strategy. Additionally, the conversion efficiency improved by 0.37%. However, the UER is decreased by 2.06%. Consequently, the overall performance was inferior to Random in Top. Furthermore, Random in Top’s inherent explicit diversity structure contributes positively to ecosystem development.

Lifelong behavior sequence. Table 4 conducted an ablation study on the impact of lifelong behavior sequence while retaining In-Context training and multiple In-Context retrieval. Without the aid of effective lifelong behavior sequence, LongRetriever struggled to effectively complete the training paradigm where all negative examples were hard negatives in In-Context training, resulting in a significant decline across all posterior metrics.

In-Context training and In-Context retrieval. Table 5 ablates the impact of In-Context training and In-Context retrieval while retaining the lifelong behavior sequence. Removing the In-Context components introduces data leakage in LongRetriever’s training. Although the absence of these components yields a significant increase in all exposure ratio and unique exposure ratio, the retrieved candidates exhibit mediocre performance in conversion efficiency.

Table 5: Ablation Study on In-Context Components

Model	AER	UER	CTR*CVR
LongRetriever	18.72%	9.29%	7.92%
w/o In-Context	38.92%	13.48%	5.61%

4 CONCLUSIONS

In this paper, LongRetriever is proposed to address the challenge of modeling more controllable and accurate diverse user interests in real industry. LongRetriever employs independent In-Context Training and Multi-Context Retrieval for each interest. LongRetriever integrates the powerful user’s lifelong behavior sequence using a search-based approach. This transforms the single fixed-length user vector into multiple precise vectors within their respective interest domains, with minimal additional training overhead. Extensive online A/B testing within our business scenarios validates LongRetriever’s effectiveness. The framework is now deployed at scale, impacting billions of users.

In future work, user information could be processed by large language models to generate optimized category assignments, thereby achieving more accurate and diverse recommendations.

References

- [1] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2020. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems* 87 (2020), 101374.
- [2] Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Yong-Nan Zhu, Zhangming Chan, Na Mou, et al. 2022. CAN: feature co-action network for click-through rate prediction. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 57–65.
- [3] Jon Nicolas Bondevik, Kwabena Ebo Bennin, Onder Babur, and Carsten Ersch. 2024. A systematic review on food recommender systems. *Expert Systems with Applications* 238 (2024), 122166.
- [4] Zheng Chai, Zhihong Chen, Chenliang Li, Rong Xiao, Houyi Li, Jiawei Wu, Jingxu Chen, and Haihong Tang. 2022. User-aware multi-interest learning for candidate matching in recommenders. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1326–1335.
- [5] Zheng Chai, Qin Ren, Xijun Xiao, Huizhi Yang, Bo Han, Sijun Zhang, Di Chen, Hui Lu, Wenlin Zhao, Lele Yu, et al. 2025. LONGER: Scaling Up Long Sequence Modeling in Industrial Recommenders. *arXiv preprint arXiv:2505.04421* (2025).
- [6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.
- [7] Jiri Hron, Karl Krauth, Michael Jordan, and Niki Kilbertus. 2021. On component interactions in two-stage recommender systems. *Advances in neural information processing systems* 34 (2021), 2744–2757.
- [8] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2553–2561.
- [9] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 2615–2623.
- [10] Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-based product retrieval in taobao search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3181–3189.
- [11] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.
- [12] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2685–2692.
- [13] Deepjyoti Roy and Mala Dutta. 2022. A systematic review and research perspective on recommender systems. *Journal of Big Data* 9, 1 (2022), 59.
- [14] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
- [15] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.