IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 February 2023, accepted 15 March 2023, date of publication 20 March 2023, date of current version 24 March 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3259462*

RESEARCH ARTICLE

# FE-TCM: Filter-Enhanced Transformer Click Model for Web Search

**YINGFEI WANG[1], JIANPING LIU[1,2], JIAN WANG[3], XIAOFENG WANG[1], MENG WANG[1], AND XINTAO CHU[1]**

[1]College of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China
[2]The Key Laboratory of Images and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China
[3]Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

Corresponding author: Jianping Liu (liujianping01@nmu.edu.cn)

This work was supported in part by the Natural Science Foundation Project of Ningxia Province, China, titled "User-Oriented Multi-Criteria Relevance Ranking Algorithm and Its Application" under Grant 2021AAC03205; in part by the Key Research and Development Program for Talent Introduction of Ningxia Province China titled "Research on Key Technologies of Scientific Data Retrieval in the Context of Open Science" under Grant 2022YCZX0009 and Grant 61862001; in part by the Starting Project of Scientific Research in the North Minzu University titled "Research of Information Retrieval Model Based on the Decision Process" under Grant 2020KYQD37; and in part by the North Minzu University Postgraduate Innovation Project under Grant YCX22178 and Grant YCX22193.

**ABSTRACT** Constructing click models and extracting implicit relevance feedback information from interaction between users and search engines are very important for improving the ranking of search results. Neural networks are effective for modeling users' click behavior, and we propose a novel Filter-Enhanced Transformer Click Model (FE-TCM) for web search. The model uses the powerful Transformer model as the backbone network for feature extraction and innovatively add a filter layer. Firstly, in order to reduce the influence of noise on user behavior data, we use the learnable filters to filter the log noise. Secondly, following the examination hypothesis, we model the attraction estimator and examination predictor respectively to output attractiveness scores and examination probabilities. A novel transformer model is used to learn the deeper representation among different features. Finally, we apply the different combination functions to integrate attractiveness scores and examination probabilities into the click prediction. From our experiments on two real-world session datasets, it is proved that FE-TCM outperforms the existing click models for the click prediction.

**INDEX TERMS** Click model, click prediction, web search, transformer.

## I. INTRODUCTION

Search result ranking is one of the major concerns in search engine researches. Click models construction which aim to improving ranking performance by using implicit relevance feedback information from the interaction between users and search engines has been paid much attention.

Traditional click models are based on the Probabilistic Graph Model (PGM) framework [1], where user behaviors are represented as a sequence of observable variables (e.g., clicks) and hidden variables (e.g., examination, skip, relevance,etc.). PGM-based click models have strong explanatory power, while this kind of model needs to manually design the dependencies among binary variables, which may over-simplify user behaviors [2]. With the development of deep learning, researchers construct click models by Neural Network (NN). NN-based click models can improve the accuracy of user behavior prediction by enhancing expression abilities and allowing flexible dependencies. Studies show that NN-based click models outperform PGM-based click models for the click prediction [2], [3], [4], [5].

Transformer [6] is the latest state of the art in sequence-to-sequence learning, and it has demonstrated outstanding performance in natural language processing tasks like text classification and context prediction [7], [8]. Some recent works have started to use Transformer to construct click models. Li et. al. [9] proposed the InterHAt model, which leverages an efficient attention aggregation strategy to learn

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.


This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

high-order feature interaction and realize click prediction. Bisht et. al. [10] developed the v-TCM model, which is also based on the Transformer architecture and learns user behaviors from vertical information.

However, user click behaviors are usually complex and noisy, and the logged data on their behavior is inherently noisy [11], [12]. Previous studies have shown that deep neural networks tend to overfit on noisy data [13], [14]. When the logged user behavior data contains noise, the performance of Transformer based on self-attention mechanism will be degraded because it pays attention to all feature embedding of sequence modeling [15].

Considering the above issues, we introduce filtering algorithms into click models and apply filtering algorithms to attenuate the noise for sequence data [16]. We suspect that when the sequence data was denoised, it will be easier to capture sequence user behaviors.

In this paper, we propose a novel Filter-Enhanced Transformer Click Model (FE-TCM) for web search. Firstly, following the examination hypothesis [17], we model attractiveness estimator and examination predictor respectively to output attractiveness scores and examination probabilities. The Transformer with multi-head self-attention mechanism is used for feature learning and extracting the dependencies between different features in user behavior sequence. Secondly, we add learnable filters between embedding layer and backbone network layer to reduce the influence of noise for sequence data. Finally, we combine attraction scores and examination probabilities through the combination layer to complete user click prediction task.

The main contributions of this paper are as follows: Firstly,we introduce filtering algorithms into the click models to reduce the influence of noise on user behavior data. Secondly, we propose a novel Filter-Enhanced Transformer Click Model (FE-TCM). FE-TCM incorporates the Transformer with multi-head self-attention mechanism for learning the position bias from user logs. Finally, From our experiments on two real-world session datasets Yandex and TREC2014, the proposed FE-TCM achieves significantly better performance than existing click models in click prediction task.

## II. RELATED WORK

### A. CLICK MODEL

In recent years, researchers have proposed numerous click models to describe users' search behaviors in search engines. Traditional click models are based on PGM framework. PGM-based click models treat user's search behaviors as a sequence of observable and hidden events, and manually design the dependencies between these binary events, which is flexible and explanatory. Craswell et. al. [18] first proposes the cascade model (CM), which assumes that users scan each search result from top to bottom until the first click. In the CM, users always leave the search engine result page (SERP) after first click and never return. However, User Browsing Model (UBM) [19], Dynamic Bayesian Network (DBN) [20], Dependent Click Model (DCM) [21] and Click Chain Model (CCM) [22] have been proposed to overcome the limitation of CM.

Due to the limited expression ability of PGM, PGM-based click models only model the important factors that affect search behaviors. If a more complex search scenario is to be considered in the PGM, the model will introduce more binary variables, and also need to manually design the dependencies among these new variables, which will cause many difficulties in the iteration and calculation of click models [23]. Therefore, Researchers try to construct click models by neural networks. Borisov et. al. [2] first proposes the neural click model (NCM). User behaviors are modeled as a sequence of vector states, which are iteratively updated with the interaction between users and search engines. Click Sequence Model (CSM) [4] follows the encoder-decoder architecture, mainly focusing on the prediction of user click sequences in search engines. Chen et. al. [3] proposes a context-aware model (CACM), It models the session context with an end-to-end neural network and jointly learns the relevance scores and the examination probability of a specific document respectively. Lin et. al. [5] proposes a graph-enhanced model (GraphCM), which combines intra-session and inter-session information by applying graph neural network and neighbor interaction technology. The experimental results show that GraphCM effectively alleviates the data sparsity and cold start problem.

Our proposed model employs the latest Transformer architecture and incorporates a filter layer remove noise from the logs. Compared with the existing click models, its performance has reached the most advanced level for the click prediction.

### B. TRANSFORMER

Transformer is the latest state of the art in sequence-to-sequence learning. It is based on a self-attention mechanism, which effectively alleviates the time dependence of RNN on long sequences. SASRec [24] and BERT4Rec [25] have proved the effectiveness of Transformer in sequence problems, so some recent works have started adapting transformer for constructing click models. Li et. al. [9] proposes the InterHAt model, which applies the efficient attention aggregation strategy to learn high-order feature interaction for the task of click prediction. Bisht and Susan [10] proposes the v-TCM model, which is based on the Transformer structure. v-TCM incorporates the vertical information type of each document in SERP (vertical bias) as an additional input to the encoder of a multi-head self-attention based transformer, apart from the query and the ranked search engine results (position bias). Zhou et. al. [26] proposes an advertising click rate prediction model SACSN based on improved Transformer structure. SACSN not only effectively models users' historical interests, but also considers the relationship with target advertisements. Chen et. al. [27] proposes BST model, which incorporates Transformer to capture the sequential signals behind users

behavior sequence, and extensive experiments prove the superiority of Transformer in modeling the user behavior sequence.

Due to the superiority of Transformer in processing sequence data, we intorduce Transformer to model users' click sequence, and learn the deeper representation of each feature by capturing the relationship with other features in the behavior sequence.

## III. PRELIMINARIES

In this section, we formulate the click model problem, and then introduce the Fourier transform.

### A. PROBLEM FORMULATION

Users click behaviors occur in each search session, and the search session $S$ can be regarded as a sequence of queries $Q_n = [q_1, q_2, \ldots, q_n]$ submitted by users. After users submit a query $q_i$ to the search engine, the search engine will return a ranked list of documents $D_i = [d_{i,1}, d_{i,2}, \ldots, d_{i,n}]$. Each document $d_{i,j}$ contains two attributes: the unique URL identifier $u_{i,j}$ and the ranking position $p_{i,j}$. The documents are ranked according to their relevance to the query and presented to users in the form of ranked list. Users examine any document and decide whether to click, if $c_{i,j} = 1$, it is regarded as a click, and 0 if not. We can define the problem of click models as follows:

Given the user's queries $Q = [q_1, q_2, \ldots, q_n]$, documents $D = [d_{1,1}, d_{1,2}, \ldots, d_{n,m}]$ and clicks $C = [c_{1,1}, c_{1,2}, \ldots, c_{n,m-1}]$, for the m-th document $d_{n,m}$ in the n-th query $q_n$ of session $S$. the task of our model is to predict whether users will click the document (i.e. the click variable $c_{n,m}$ ).

In this paper, we model examination predictor and attractiveness estimator respectively to output attractiveness scores and examination probabilities. In the attractiveness estimator, we take user queries $q$, documents $d$, click variables $c$ and ranked position $p$ as inputs. In the examination predictor, considering that the user's examination operation is only affected by his/her operation on the previous results in the current query, we take click variables $c$ and ranked position $p$ as inputs. Finally, we combine the attraction score $\mathcal{A}_{n,m}$ and examination probability $\mathcal{E}_{n,m}$ to output the final click probability.

### B. FOURIER TRANSFORM AS FILTER LAYER

Discrete Fourier Transform (DFT) is essential in the digital signal processing field [28]. DFT is a discrete form of continuous Fourier Transform in both time domain and frequency domain. In practical applications, the Fast Fourier Transform (FFT) is usually used to efficiently calculate DFT [29]. Given a sequence $\{x_k\}$ with $k \in [0, N-1]$, the sequence is converted into frequency domain by 1D DFT:

$$x_n = \sum_{k=0}^{N-1} x_k e^{-i\frac{2\pi}{N}kn}, n = 0, 1, \ldots, N-1 \tag{1}$$

Given the $DFTx_n$, We can convert it into the original sequence by inverse DFT (IDFT):

$$x_k = \frac{1}{N} x_n e^{\frac{2\pi i}{N}nk} \tag{2}$$

Because FFT can transform the input signal into frequency domain Where periodic features are easier to capture, they are widely used in the digital signal processing field to filter noise signals. We consider using FFT to reduce the influence of noise features in user sequence data.

## IV. MODEL FRAMEWORK

As shown in Figure 1, we will introduce the overall framework of FE-TCM.

### A. EMBEDDING LAYER

The first component is the embedding layer, which embeds all input features into fixed-size low-dimensional vectors. FE-TCM takes query $q$, document $d$, user click $c$ and ranked position $p$ as inputs, these original ID features are transformed into a high-dimensional sparse features via one-hot encoding, then we convert high-dimensional sparse vector into low-dimensional dense vector through embedding layer:

$$v_q = Emb_q(q), v_d = Emb_d(d), v_c = Emb_c(c), v_p = Emb_p(p) \tag{3}$$

where $Emb_* \in \mathbb{R}^{N_* \times l_*}, * \in \{q, d, c, p\}$. $N_*$ and $l_*$ denote the input feature size and embedding size.

### B. FILTER LAYER

As shown in Figure 2, based on the embedding layer, we develop the input features of transformer encoder by stacking multiple learnable filter blocks. In the filtering layer, we first perform filtering operation for each dimension of features in the frequency domain, in order to avoid over-fitting of FE-TCM, we perform skip connection and layer normalization. Given the representation matrix $F^l \in \mathbb{R}^{n \times d}$, we first perform FFT along its dimension to transform it into the frequency domain:

$$X^l = \mathcal{F}(F^l) \tag{4}$$

where $\mathcal{F}(\cdot)$ represents one-dimensional FFT, and $X^l$ represents frequency spectrum of $F^l$. Then, We can modulate the spectrum by multiplying a learnable filter $W = \mathbb{C}^{n \times d}$:

$$\widetilde{X}^l = W \odot X^l \tag{5}$$

Finally, we adopt the inverse FFT to transform the modulated spectrum back to the time domain and update the sequence representations:

$$\widetilde{F}^l \leftarrow F^{-1}(\widetilde{X}^l) \tag{6}$$

where $F^{-1}(\cdot)$ denotes the inverse 1D FFT, which converts the complex tensor into a real number tensor. Through FFT and inverse FFT operations, the noise of recorded data can be effectively reduced, thus obtaining more pure feature embedding.

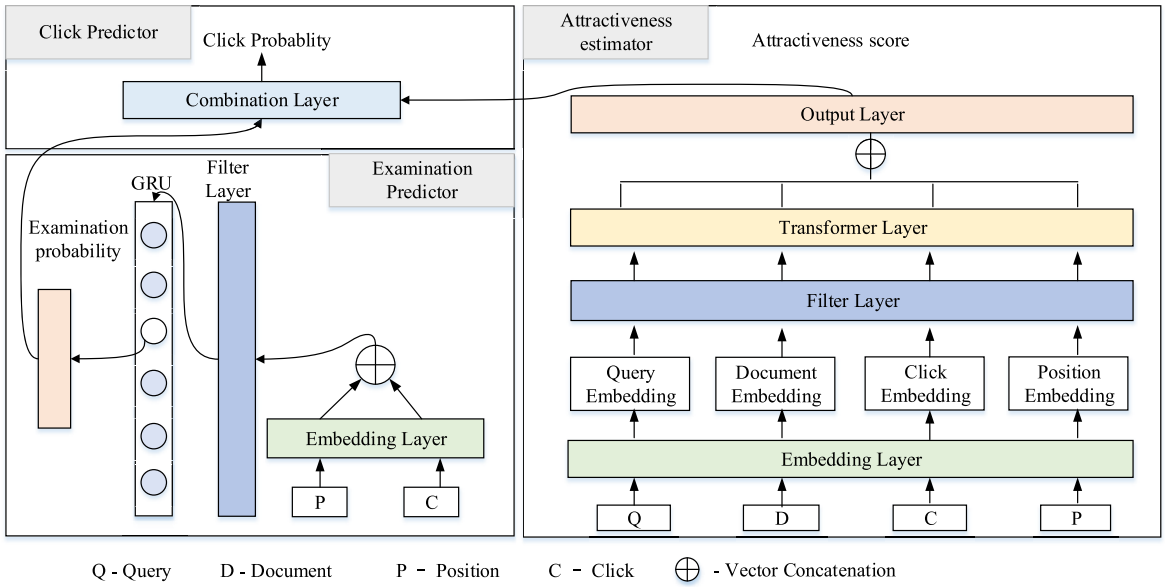



**FIGURE 1.** Overall framework of FE-TCM. FE-TCM consists of an attractiveness estimator and an examination predictor. The attractiveness predictor is used to estimate the attraction score $\mathcal{A}_{n,m}$. The examination predictor is used to predict the examination probability $\mathcal{E}_{n,m}$. FE-TCM integrates $\mathcal{E}_{n,m}$ and $\mathcal{A}_{n,m}$ through a combination layer to predict user click behaviors.

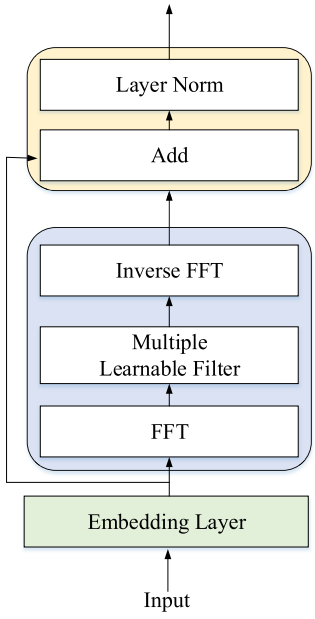



**FIGURE 2.** The overall framework of the filter layer.

In order to alleviate the problem of gradient vanishing and unstable training problems, we also incorporate the skip connection, layer normalization and dropout operations. LayerNorm represents layer normalization, which is mainly used to speed up the convergence of the model; Dropout is a random inactivation, which is used to prevent over-fitting, and the skip connection is adopted to reduce the learning load of the model. The formula is as follows:

$$\widetilde{F}^l = LayerNorm(F^l + Dropout(\widetilde{F}^l)) \tag{7}$$

### C. EXAMINATION PREDICTOR

Following the examination hypothesis [17], the probability that users click the document depends on the user examining the document and considering it relevant to the query. The examination predictor aims to predict whether the user will continue to examine the document based on his/her session context. In FE-TCM, we follow the hypothesis of previous work [3], [5], the user's examination behavior is only affected by his/her operation on previous documents. Therefore, for the current document $d_{i,j}$, we apply a session-level GRU [30] to encode the ranking position $p$ and historical clicks $c$ in the same session.

$$x_{i,j} = [v_{pi,j} \oplus v_{ci,j}] \tag{8}$$

$$\mathcal{E}'_{i,j} = GRU(x_{1,1}, \ldots, x_{i,j}) \tag{9}$$

Finally, we apply a linear layer and a sigmoid function to realize normalization and output the final examination probability $\mathcal{E}$.

$$\mathcal{E}_{i,j} = Sigmoid(Linear(\mathcal{E}'_{i,j})) \tag{10}$$

### D. ATTRACTIVENESS ESTIMATOR

Next, we will introduce the attractiveness estimator. The attractiveness estimator aims to estimate the attractiveness of each document $d_{i,j}$ to the user who issues the query $q_i$. After the filter layer, FE-TCM embeds all the features as inputs of Transformer encoder. Transformer learns the deeper representation of each feature by capturing the relationship with other features in the behavior sequence.

#### 1) MULTI-HEAD SELF-ATTENTION LAYER

The embedding vector $E$ is calculated by multi-head self-attention, and the process is as follows: $E$ is converted into

query vector $Q = EW_i^Q$, keyword vector $K = EW_i^K$ and value vector $V = EW_i^V$ by $W_i^Q$, $W_i^K$ and $W_i^V$ respectively, where $W_i^Q$, $W_i^K$ and $W_i^V$ are parameter matrices. Then we compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$,and apply a softmax function to obtain the weights on the values.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{11}$$

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions.Following [6], we use the multi-head attention:

$$head_i = Attention(EW_i^Q, EW_i^K, EW_i^V) \tag{12}$$

$$S = concat(head_1, \ldots, head_h)W^O \tag{13}$$

where $head_i$ represents the i-th self-attention. In our task, the self-attention operation takes the filtered embedding of query vector, document vector, ranked position vector and target click vector as inputs, then converts them into three matrices through linear projection, and feeds them to the attention layer.

#### 2) POINT-WISE FEED-FORWARD NETWORKS

The essence of multi-head attention is linear transformation, we incorporate Point-wise Feed-Forward Networks (FFN) to further enhance the model with non-linearity, which is defined as follows.

$$F = FFN(S) = ReLU(SW^{(1)} + b^{(1)})W^{(2)} + b^{(2)} \tag{14}$$

where $F$ is the output vector of the FFN, and $W^{(1)}$,$W^{(2)}$, $b^{(1)}$,$b^{(2)}$ are learnable parameters.

Finally, we concatenate the query vector $q$, the document vector $d$, the click vector $c$ and the position vector $p$ output by Transformer, and generate the final attraction scores $\mathcal{A}$ through a linear layer and a sigmoid function.

$$output = [q \oplus d \oplus c \oplus p] \tag{15}$$

$$\mathcal{A} = Sigmoid(Linear(output)) \tag{16}$$

### E. COMBINATION LAYER

The click prediction module combines the examination probability $\mathcal{E}$ and attraction score $\mathcal{A}$ to output the final click probability. We have implemented five different combination functions, as shown in Table 1.

For *mul* function, we multiply the attraction score $\mathcal{A}$ with the examination probability $\mathcal{E}$ directly. The *exp_mul* follows the examination hypothesis and increases the model capacity by adding learnable parameters. The *linear* function and *nonlinear* function further discuss the relationship between $\mathcal{E}$ and $\mathcal{A}$. The *sigmoid_log* function uses sigmoid function and logarithmic function, and finally obtain a simple formula. Among the five combination functions, only *mul*, *exp_mul* and *sigmoid_log* support the examination hypothesis.

**TABLE 1. Combination functions.**

| Function | Forluma | Support E.H.? |
|---|---|---|
| mul | $\mathcal{C} = \mathcal{A} \cdot \mathcal{E}$ | Yes |
| exp_mul | $\mathcal{C} = \mathcal{A}^\lambda \cdot \mathcal{E}^\mu$ | Yes |
| sigmoid_log | $\mathcal{C} = 4\sigma(log(\mathcal{A}) \cdot \sigma(log(\mathcal{E}))$ | Yes |
| linear | $\mathcal{C} = \alpha \cdot \mathcal{A} + \beta \cdot \mathcal{E}$ | No |
| nonlinear | $\mathcal{C} = MLP(\mathcal{A}, \mathcal{E})$ | No |

**TABLE 2. Configuration of FE-TCM.**

| Configuration of FE-TCM | | | |
|---|---|---|---|
| embedding size | 64 | batchsize | 64 |
| hidden size | 64 | learning rate | 0.001 |
| head number | 8 | dropout | 0.5 |
| Transformer block | 1 | weight decay | $10^{-5}$ |

**TABLE 3. The dataset statistics.**

| Dataset | Session | Query | Search Engine |
|---|---|---|---|
| Yandex | 200, 000 | 376, 965 | Yandex |
| TREC2014 | 1257 | 5443 | Irdia |

### F. LOSS FUNCTION

In order to learn the weights and parameters of the model, We take the cross entropy loss function as the objective function, and the formula is expressed as:

$$L = -\frac{1}{N}\sum_i\sum_j(C_{i,j}logP_{i,j} + (1 - (C_{i,j})log(1 - P_{i,j})) \tag{17}$$

where $n$ is the number of training samples. $C_{i,j}$ and $P_{i,j}$ denote the true click signal and the predicted click probability of the r-th result in the i-th query session in the testing set.

## V. EXPERIMENT

In this section, we conduct experiments to answer the following questions:

**RQ1** Compared with baseline click models, Does FE-TCM achieve the best performance in click prediction task on Yandex and TREC2014 two datasets?

**RQ2** Which combination function performs best when integrating attraction scores and examination probabilities?

**RQ3** What is the influence of different components in FE-TCM?

**RQ4** Does the learnable filters improve the model performance?

### A. EXPERIMENTAL SETUP

#### 1) IMPLEMENTATION DETAILS

The operating system used is Windows, the GPU is NVIDIA TITAN V. We train our model with Adam optimizer. We give detailed model parameters in Table 2.

#### 2) DATASET

We conducted experiments on two public session datasets. The statistics of the datasets can be found in Table 3.

(1)Yandex[1]: The dataset includes user sessions extracted from Yandex logs, with user ids, queries, query terms, URLs, their domains, URL rankings and clicks.

[1]https://www.kaggle.com/c/yandex-personalized-web-search-challenge

(2)TREC2014[2]: The dataset includes queries, URLs, URL ranking position, clicks, and the time spent by users reading the web page corresponding to each click.

Due to the limitation of memory, we randomly sample the sessions in the Yandex dataset. All datasets are divided into training set, validation set and test set according to the ratio of 8:1:1.

#### 3) EVALUTION METRIC

We compare our proposed model with other baseline models in the click prediction, and we report the perplexity (PPL) and log-likelihood (LL) of each model in the paper. Perplexity and log-likelihood are defined as follows:

$$PPL@r = 2^{-\frac{1}{N}\sum_{i=1}^{N} C_{i,r} log P_{i,r} + (1 - C_{i,r}) log(1 - P_{i,r})} \tag{18}$$

$$LL = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} C_{i,j} log P_{i,j} + (1 - C_{i,j}) log(1 - P_{i,j}) \tag{19}$$

where $r$ represents the ranking position on the search engine results page, $N$ represents the number of sessions, $M$ represents the number of results in a query, $C_{i,r}$ is the actual click, and $P_{i,r}$ is the predicted click probability of the r-th document of the i-th query in the test set. We average the perplexity values of all positions to get the overall perplexities of the model. The lower value of perplexity and the higher value of log-likelihood correspond to the better click prediction performance.

#### 4) BASELINES

The existing click models can be divided into two class: PGM-based click models and NN-based click models [31]. For PGM-based click models, we choose five representative models: UBM, DCM, DBN, SDBN and CCM, all of which open-source implementations are available.[3] For NN-based click models, we choose NCM and GraphCM as baseline models.

### B. PERFORMANCE COMPARISON (ANSWER RQ1)

We perform click prediction task for each click model to compare the performance. The results are shown in Table 4, from which we can get the following observations:

(1)Among all PGM-based click models, UBM has the advantages of simple structure, low computational cost, and always achieves good experimental results, so it performs best in all PGM-based click models.

(2)NCM, GraphCM and FE-TCM are superior to all PGM-based click models in click prediction task. GraphCM performs best among all baseline models. GraphCM combines intra-session and inter-session information by using graph neural network (GAT) and neighbor interaction techniques, so GraphCM can better capture more subtle patterns in user click behaviors.

(3)Our proposed FE-TCM obviously outperforms all baseline models. This improvement proves the superiority of Transformer with multi-head self-attention mechanism in processing sequence data, and the effectiveness of the learnable filters in alleviating the impact of noise in click models.

**TABLE 4. Overall performance of each click model. The best results are shown in bold. The lower value of PPL and the higher value of LL correspond to the better click prediction performance.**

| Model | Yandex | | TREC2014 | |
|---|---|---|---|---|
| | LL | PPL | LL | PPL |
| CCM | -0.2975 | 1.3393 | -0.3632 | 1.3081 |
| DCM | -0.3100 | 1.3480 | -0.3728 | 1.2702 |
| DBN | -0.2980 | 1.3354 | -0.3633 | 1.2847 |
| SDBN | -0.3034 | 1.3353 | -0.3707 | 1.2858 |
| UBM | -0.2530 | 1.3320 | -0.1732 | 1.2236 |
| NCM | -0.2333 | 1.2851 | -0.1652 | 1.1880 |
| GraphCM | -0.2192 | 1.2652 | -0.1529 | 1.1721 |
| FE-TCM | **-0.2106** | **1.2543** | **-0.1475** | **1.1689** |

**TABLE 5. Performance comparison of models with different combination functions.**

| Combination function | Yandex | | TREC2014 | |
|---|---|---|---|---|
| | LL | PPL | LL | PPL |
| FE-TCM *linear* | -0.2328 | 1.2845 | -0.1672 | 1.1980 |
| FE-TCM *nonlinear* | -0.2410 | 1.2917 | -0.1754 | 1.2046 |
| FE-TCM*sigmoid_log* | -0.2199 | 1.2675 | -0.1494 | 1.1712 |
| FE-TCM *mul* | -0.2179 | 1.2654 | -0.1487 | 1.1706 |
| FE-TCM *exp_mul* | **-0.2106** | **1.2543** | **-0.1475** | **1.1689** |

### C. COMBINATION FUNCTION (ANSWER RQ2)

We study the influence of different combination functions on the performance of FE-TCM. From Table 5, we can obtain that *exp_mul* function can achieve the best results in the combination layer. Compared with *mul*, *exp_mul* and *sigmoid_log* functions, the *linear* and *nonlinear* functions have relatively poor results because they do not support the examination hypothesis.

### D. ABLATION STUDY (ANSWER RQ3, RQ4)

In order to study the contribution of each module to the overall performance of FE-TCM, and to verify whether the learnable filters can improve the performance of the model, we conducted several comparative experiments. The results are presented in Table 6.

It can be seen from Table 6, compared with long short-term memory (LSTM), the LL value and PPL value can be increased by about 0.1% by using gated recurrent unit (GRU) in the examination predictor; Besides, by adding filter layers to the examination predictor and attraction estimator, LL and PPL are both improved, which indicates that the learnable filters can improve the performance of GRU and Transformer structures. When the filter layer is added to both examination predictor and attraction estimator, the overall PPL value of the model is increased by 1.8% on Yandex dataset, and the PPL value is 1.2543, LL value is increased by 1.3%, and LL value

[2]https://trec.nist.gov/data/session2014.html
[3]https://github.com/markovi/PyClick

**TABLE 6. Analysis of key components of FE-TCM.**

| Method | Yandex | | TREC2014 | |
|---|---|---|---|---|
| | LL | PPL | LL | PPL |
| Transformer+LSTM | -0.2238 | 1.2736 | -0.1498 | 1.1718 |
| Transformer+GRU | -0.2229 | 1.2725 | -0.1488 | 1.1706 |
| Transformer+Attraction+Examination_Filter+GRU | -0.2218 | 1.2658 | -0.1496 | 1.1714 |
| Transformer+Attraction_Filter+Examination+GRU | -0.2228 | 1.2721 | -0.1477 | 1.1691 |
| Transformer+Attraction_Filter+Examination_Filter+GRU | -0.2106 | 1.2543 | -0.1475 | 1.1689 |

is −0.2106; For the TREC2014 dataset, the overall PPL value of the model increased by 0.17% and the PPL value is 1.1689, the LL value increased by 0.13%, and the LL value is -0.1475. The results prove the effectiveness of adding learnable filters between the embedded layer and backbone network layers.

## VI. DISCUSSION AND FUTURE WORK

### A. DISCUSSION

In this paper, we propose a novel Filter-Enhanced Transformer Click Model (FE-TCM) for web search. FE-TCM consists of an attraction estimator and an examination predictor for click prediction. Specifically the attraction estimator incorporates the Transformer architecture to extract the dependence of different features in users behavior sequence. To achieve the best performance, we further study five combination functions to integrate attraction and examination into click prediction.

We conducted extensive experiments on two real-world session datasets to answer four research questions, and found that: 1) Our proposed model outperforms existing advanced click models in the click prediction task due to the use of the Transformer architecture and learnable filters. 2) Transformer with multi-head self-attention mechanism is more suitable for capturing sequence signals, and can better learn the position bias from user logs. This structure can be used as the most effective feature extraction network for click prediction. 3) Click models are mainly based on logged user behavior data to fit model parameters, which usually contains noisy interactions. Ablation study shows that filtering algorithms from digital signal processing can effectively alleviate the influence of noise in click models, and proves that adding learnable filters between the embedding layer and the backbone network layer can significantly improve the performance of the model. 4) Among the five combination functions, *exp_mul* function has the best comprehensive performance because it supports the examination hypothesis and has more learnable parameters to model user behaviors flexibly.

### B. FUTURE WORK

Through these experiments, we fully understand the advantages and limitations of FE-TCM, and these limitations can further inspire some future work. For example, 1) For input features, we plan to expand click models by combining rich contextual information (e.g., dwell time, action type, vertical information, etc.), and making full use of more user behavior information (e.g., visual bias, mouse movement, etc.). 2) Many existing studies have shown that incorporating the embedding learned from graph structure can significantly improve model performance [3], [5], [32]. Graph structure must be exploited in the future to further improving session and click modeling. 3) In the users' intention prediction, existing studies incorporate the pre-training method to calculate the topic relevance and assist intention prediction [33]. Because this type of data contains query and document information, integrating topic relevance into click models can also be used as one of the directions to accurately predict users' click behaviors.

## VII. CONCLUSION

In this paper, we propose a new Filter-Enhanced Transformer Click Model (FE-TCM) for web search. We introduce the filtering algorithms borrowed from signal processing field, and ablation experiments prove that the learnable filters can effectively alleviate the influence of noise in sequence data. In addition, we model an attraction predictor and a examination predictor respectively, and apply Transformer to extract the dependence of different features in users behavior sequence. We have conducted extensive experiments on two real-world session datasets, and proves the superiority of our proposed model in click prediction task compared with the existing advanced click models.

## REFERENCES

[1] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[2] A. Borisov, I. Markov, M. De Rijke, and P. Serdyukov, "A neural click model for web search," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 531–541.

[3] J. Chen, J. Mao, Y. Liu, M. Zhang, and S. Ma, "A context-aware click model for web search," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 88–96.

[4] A. Borisov, M. Wardenaar, I. Markov, and M. De Rijke, "A click sequence model for web search," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jun. 2018, pp. 45–54.

[5] J. Lin, W. Liu, X. Dai, W. Zhang, S. Li, R. Tang, X. He, J. Hao, and Y. Yu, "A graph-enhanced click model for web search," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2021, pp. 1259–1268.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, K. Aiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[7] P. Li, P. Zhong, K. Mao, D. Wang, X. Yang, Y. Liu, J. Yin, and S. See, "ACT: An attentive convolutional transformer for efficient text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 13261–13269.

[8] N. Wang, W. Zhou, J. Wang, and H. Li, "Transformer meets tracker: Exploiting temporal context for robust visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1571–1580.

[9] Z. Li, W. Cheng, Y. Chen, H. Chen, and W. Wang, "Interpretable click-through rate prediction through hierarchical attention," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 313–321.

[10] K. Bisht and S. Susan, "V-TCM: Vertical-aware transformer click model for web search," in *Proc. 37th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2022, pp. 1917–1920.

[11] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2006, pp. 19–26.

[12] A. Said, B. J. Jain, S. Narr, and T. Plumbaum, "Users and noise: The magic barrier of recommender systems," in *Proc. Int. Conf. User Model., Adaptation, Personalization*. Cham, Switzerland: Springer, 2012, pp. 237–248.

[13] R. Caruana, S. Lawrence, and C. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 1–7.

[14] J. Lever, M. Krzywinski, and N. Altman, "Points of significance: Model selection and overfitting," *Nature Methods*, vol. 13, no. 9, pp. 703–705, 2016.

[15] K. Zhou, H. Yu, W. X. Zhao, and J.-R. Wen, "Filter-enhanced MLP is all you need for sequential recommendation," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2388–2399.

[16] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1975.

[17] A. Chuklin, I. Markov, and M. D. Rijke, "Click models for web search," *Synth. Lect. Inf. Concepts, Retr., Services*, vol. 7, no. 3, pp. 1–115, 2015.

[18] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *Proc. Int. Conf. Web Search Web Data Mining*, 2008, pp. 87–94.

[19] G. E. Dupret and B. Piwowarski, "A user browsing model to predict search engine click data from past observations," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2008, pp. 331–338.

[20] O. Chapelle and Y. Zhang, "A dynamic Bayesian network click model for web search ranking," in *Proc. 18th Int. Conf. World Wide Web*, Apr. 2009, pp. 1–10.

[21] F. Guo, C. Liu, and Y. M. Wang, "Efficient multiple-click models in web search," in *Proc. 2nd ACM Int. Conf. Web Search Data Mining*, Feb. 2009, pp. 124–131.

[22] F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos, "Click chain model in web search," in *Proc. 18th Int. Conf. World Wide Web*, Apr. 2009, pp. 11–20.

[23] Y. Liu, C. Wang, M. Zhang, and S. Ma, "User behavior modeling for better web search ranking," *Front. Comput. Sci.*, vol. 11, no. 6, pp. 923–936, Dec. 2017.

[24] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 197–206.

[25] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag.*, Nov. 2019, pp. 1441–1450.

[26] X. H. Z. Fei, "Improved transformer based model for click-through rate prediction," *Appl. Res. Comput.*, vol. 38, no. 8, pp. 1–5, Aug. 2021.

[27] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, "Behavior sequence transformer for e-commerce recommendation in Alibaba," in *Proc. 1st Int. Workshop Deep Learn. Pract. High-Dimensional Sparse Data*, Aug. 2019, pp. 1–4.

[28] S. S. Soliman and M. D. Srinath, *Continuous and Discrete Signals and Systems*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1990.

[29] C. V. Loan, *Computational Frameworks for the Fast Fourier Transform*. Philadelphia, PA, USA: SIAM, 1992.

[30] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[31] D. B. Bakhtiarvand and S. Farzi, "An ensemble click model for web document ranking," *Int. J. Eng.*, vol. 33, no. 7, pp. 1208–1213, 2020.

[32] J.-Y. Jiang and W. Wang, "RIN: Reformulation inference network for context-aware query suggestion," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2018, pp. 197–206.

[33] X. Zuo, Z. Dou, and J.-R. Wen, "Improving session search by modeling multi-granularity historical query change," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Feb. 2022, pp. 1534–1542.

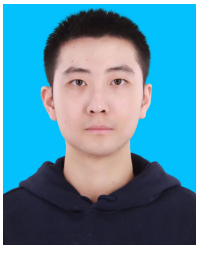

**YINGFEI WANG** is currently pursuing the master's degree with the College of Computer Science and Engineering, North Minzu University. Her research interests include interactive information retrieval and click model. Her research has been published in the International Conference on Cloud Computing and Intelligent Systems (CCIS 2022) and *Scanning*, in 2023.

**JIANPING LIU** received the Ph.D. degree in information technology and digital agriculture from the Chinese Academy of Agricultural Sciences. He is currently a Lecturer in information sciences with the College of Computer Science and Engineering, North Minzu University. His research has been published in *Library & Information Science Research*, in 2019, *Data Science Journal*, in 2020, and *Scanning*, in 2023.

**JIAN WANG** received the Ph.D. degree in geoinformatics from the Chinese Academy of Sciences. He is currently a Professor in information sciences with the Agricultural Information Institute, Chinese Academy of Agricultural Sciences. His research was published in *Sensor Letters*, in 2010, *Data Science Journal*, in 2020, and *Library & Information Science Research*, in 2019.

**XIAOFENG WANG** received the Ph.D. degree in computer software and theory from Guizhou University. He is currently an Associate Professor in information sciences with the College of Computer Science and Engineering, North Minzu University. His research was published in IEEE TRANSACTIONS ON TOPICS EMERGING IN COMPUTING, in 2019, and *Journal of Software*, in 2021.

**MENG WANG** is currently pursuing the master's degree with the College of Computer Science and Engineering, North Minzu University. Her research interest includes interactive information retrieval. Her research has been published in International Conference on Cloud Computing and Intelligent Systems (CCIS 2022) and *Scanning*, in 2023.



**XINTAO CHU** is currently pursuing the master's degree with the College of Computer Science and Engineering, North Minzu University. His research interests include interactive information retrieval and deep learning. His research has been published in International Conference on Cloud Computing and Intelligent Systems (CCIS 2022) and *Scanning*, in 2023.