CSPM: A Contrastive Spatiotemporal Preference Model for CTR Prediction in On-Demand Food Delivery Services

Guyu Jiang*, Xiaoyun Li*, Rongrong Jing, Ruoqi Zhao, Xingliang Ni[†], Guodong Cao, Ning Hu

Alibaba Group

Beijing&Shanghai, China

{guyu.qgy,xiaoyun.li,rongrongjing.jrr,ruoqi.zrq,wangzhe.nxl,guodong.cao,huning.hu}@alibaba-inc.com

ABSTRACT

Click-through rate (CTR) prediction is a crucial task in the context of an online on-demand food delivery (OFD) platform for precisely estimate the probability of a user clicking on food items. Unlike universal e-commerce platforms such as Taobao and Amazon, user behaviors and interests on the OFD platform are more location and time-sensitive due to limited delivery ranges and regional commodity supplies. However, existing CTR prediction algorithms in OFD scenarios concentrate on capturing interest from historical behavior sequences, which fails to effectively model the complex spatiotemporal information within features, leading to poor performance. To addresses this challenge, this paper introduces the Contrastive Spatiotemporal Preference Model (CSPM), which disentangles users' spatiotemporal preferences from multiple-field features under different search states using three modules: contrastive spatiotemporal representation learning (CSRL), spatiotemporal preference extractor (StPE), and spatiotemporal information filter (StIF) . CSRL utilizes a contrastive learning framework to generate a spatiotemporal activation representation (SAR) for the search action. StPE employs SAR to activate users' diverse preferences related to location and time from the historical behavior sequence field, using a multi-head attention mechanism. StIF incorporates SAR into a gating network to automatically capture important features with latent spatiotemporal effects. Extensive experiments conducted on two large-scale industrial datasets demonstrate the state-of-the-art performance of CSPM. Notably, CSPM has been successfully deployed in Alibaba's online OFD platform Ele.me, resulting in a significant 0.88% lift in CTR, which has substantial business implications.

CCS CONCEPTS

• Information systems; • Information retrieval; • Retrieval models and ranking; • Learning to rank;

KEYWORDS

Click-through Rate Prediction, On-demand Food Delivery, Spatiotemporal Preference Modeling

CIKM'23, October 10–21, 2023, Birmingham, UK

© 2023 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-x/YY/MM...\$15.00 https://doi.org/XX.XX

ACM Reference Format:

Guyu Jiang^{*}, Xiaoyun Li^{*}, Rongrong Jing, Ruoqi Zhao, Xingliang Ni[†], Guodong Cao, Ning Hu. 2023. CSPM: A Contrastive Spatiotemporal Preference Model for CTR Prediction in On-Demand Food Delivery Services. In Proceedings of 32nd ACM Conference on Information and Knowledge Management (CIKM'23). ACM, New York, NY, USA, 5 pages. https://doi.org/XX.XX

1 INTRODUCTION

With the advancements in Location-Based Services (LBS) technology, on-demand food delivery (OFD) platforms like Uber Eats, Door Dash, Meituan-Dianping, and Ele.me have witnessed substantial growth in recent years, delivering millions of meals worldwide on a daily basis. Click-through rate (CTR) prediction, a crucial component of OFD platforms' advertising and recommendation systems, aims to predict the likelihood of user interaction with recommended items. For example, in Ele.me's retrieval system, users simply provide their addresses and enter food keywords, and they are immediately presented with the most relevant candidate shops and commodities for selection. Once users make their desired purchases through the payment page, timely delivery is arranged by the delivery driver.

Unlike universal e-commerce platforms such as Amazon and Taobao, user click behaviors in OFD platforms are more sensitive to factors like location and time, necessitating corresponding advertising systems to account for users' spatiotemporal preferences. Numerous methods [3, 4, 8, 9, 12] have been proposed for user behavior modeling, exhibiting promising experimental results. For instance, StEN [8] explicitly models personalized spatiotemporal preferences by incorporating historical behaviors using a complex activation and attention-based network. CaledarGNN [12] captures time structures from spatiotemporal user behaviors but overlooks the sensitive spatial needs in food delivery. BASM [4] introduces a bottom-up adaptive model that leverages embedding, semantic, and classification layers to learn the spatiotemporal data distribution and enhance model fitting capacity. TRISAN [9] integrates the temporal relatedness of a user's historical click sequences to model the geographic closeness of items and requests based on distance or semantic similarity.

However, existing methods solely focus on capturing spatiotemporal interest from historical behavior sequences, which fails to fully exploit the spatiotemporal information contained in advertising system features. In fact, the spatiotemporal information embedded within these features is intricate in OFD scenarios. Alongside historical behavior-derived spatiotemporal interest, empirical observations for the inherent attribute in LBS systems suggest that seemingly unrelated features possess latent effects that contribute to modeling user spatiotemporal preferences. For example, while

^{*}These authors contributed equally to this work. † Corresponding author.

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 ACM 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.

subsidy releases of red packets are generally not spatiotemporal features, they exhibit spatiotemporal characteristics in practice due to specific business strategies, attracting users who appreciate discounts. Moreover, spatiotemporal preferences can be influenced by the current search state, including the query, user location, and time information, further increasing the modeling challenge. For instance, when users search for "beef", they may prefer to purchase raw food for cooking if they are at home, whereas fast food may be preferred if they are at work.

To address these challenges, we propose the Contrastive Spatiotemporal Preference Model (CSPM) that considers both explicit spatiotemporal preferences relevant to users' historical behaviors and implicit spatiotemporal preferences contained in context and item features. CSPM utilizes the contrastive spatiotemporal representation learning (CSRL) module to transform user search behavior, including the query, time, and location features, into a spatiotemporal activation representation (SAR) that will be used to model the subsequent spatiotemporal preferences. To comprehensively capture the spatiotemporal preference patterns embedded in multi-field features, CSPM introduces the spatiotemporal preference extractor (StPE) module and the spatiotemporal information filter (StIF) module. StPE module employs a multi-head attention mechanism and SAR from CSRL to extract explicit spatiotemporal preferences correlated with users' historical behaviors. StIF module incorporates SAR into a gating network to automatically capture important features with latent spatiotemporal effects. By jointly optimizing the CSRL triplet loss and the CTR prediction cross-entropy loss, CSPM learns a comprehensive spatiotemporal preference model to enhance CTR prediction accuracy. The contributions of this paper can be summarized as follows:

- We propose the contrastive spatiotemporal preference model, which utilizes contrastive spatiotemporal representation learning to generate an activation signal for the current search state, enabling the modeling of explicit and implicit spatiotemporal preferences from multi-field features.
- We develop two user preference model modules: the spatiotemporal preference extractor, which extracts explicit spatiotemporal preferences correlated with users' historical behaviors, and the spatiotemporal information filter, which activates implicit spatiotemporal information contained in context and item features.
- We evaluate CSPM on two large-scale real-world datasets, and the experimental results demonstrate its significant outperformance compared to other state-of-the-art methods in CTR prediction. Additionally, we release an industrial dataset for the OFD industry to address the scarcity of public datasets in this domain.

2 METHODOLOGY

2.1 Preliminary

Considering a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where x_i and y_i represent the feature set and binary click label of the i_{th} sample. In an OFD recommendation system, the feature set usually can typically be divided into several fields, denoted as $x_i = (x_i^e), e \in$

 $\{E_S, E_I, E_Q, E_L, E_T, E_U, E_C\}$. These superscripts correspond to different fields, including user behavior sequence, candidate item from the recall stage, search query, search location, search time, user, and context. We formulate the task of CTR prediction as a learning problem, where the goal is to estimate the probability

$$\hat{y} = \mathbb{P}(y = 1 | x, \mathcal{D}) = f(x).$$
(1)

Here, $f(\cdot)$ represents a learnable function, and $\hat{y} \in [0, 1]$ represents the predicted probability of the user clicking on the target item. To compute the prediction error of CTR, we employ the cross-entropy loss given by:

$$\mathcal{L}_{CTR} = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i).$$
(2)

2.2 Contrastive Spatiotemporal Representation Learning

The user's search action under different spatiotemporal contexts, could imply distinct needs and provides valuable insights into extracting historical and latent spatiotemporal preferences. In this paper, we utilize query, location, and time features x^e ($e \in \{E_O, E_L, e_N\}$ E_T }) to describe the search state or action. To enhance the distinguishability of the search state and activate subsequent spatiotemporal preferences effectively, we introduce a CSRL module. This module assumes that representations of semantically similar search queries under close geohashes and within the same time period should be close together in the representation space, while dissimilar search states should be far apart. CSRL transforms these search state features into low-dimensional, dense, continuous vectors using embedding techniques and utilizes a DCN-v2 [14] to generate a high-order cross-field vector s, called spatiotemporal activation representation (SAR). For constructing positive and negative pairs, we select locations from the same batch of recall results and assign temporal proximity by considering a half-hour interval ahead or behind the search time. Positive pairs are formed with these spatiotemporal constraints, while negative pairs are randomly selected. The contrastive loss in CSRL adopts the triplet loss as follows:

$$\mathcal{L}_{CL} = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{n_v} \max(\cos(s_i^p, s_i) - \cos(s_i^{n_j}, s_i) + m, 0), \quad (3)$$

where n_v represents the number of randomly selected negative samples, $\cos(\cdot, \cdot)$ calculates the cosine similarity between the two vectors, *m* is a margin, and s_i^p and $s_i^{n_j}$ denote the SARs of positive and negative samples, respectively.

2.3 Spatiotemporal Preference Extractor

The behavior sequence records a user's combination of purchase incidents within a specific time period. We argue that the current search action plays a pivotal role in eliciting a user's historical preference. Therefore, we propose the StPE module, which leverages the SAR of search state from CSRL to query the behavior sequence. Drawing on the ideas of [10], the backbone of StPE is a multi-head attention mechanism [11] that concatenates the spatiotemporal activation representation and the candidate item embedding as the activation query. The user behavior sequence is regard as the key and value by different linear transformations. The formulation of CSPM: A Contrastive Spatiotemporal Preference Model for CTR Prediction in On-Demand Food Delivery Services

CIKM'23, October 10-21, 2023, Birmingham, UK



Figure 1: The framework of our proposed CSPM. CSPM contains three main modules: CSRL, StPE, and StIF.

StPE can be expressed as follows:

$$u_{h} = \text{Attention}\left(\left[v^{E_{C}};s\right]W_{h}^{Q}, U^{E_{S}}W_{h}^{K}, U^{E_{S}}W_{h}^{V}\right), \qquad (4)$$

where v^{E_C} and U^{E_S} denote the candidate item-field representation and the user sequence-field representation matrix, respectively. Attention(·) represents the self-attention mechanism, $[\cdot, \cdot]$ is the concatenation operator, and W_h^Q , W_h^K , and W_h^V are the parameters of the h_{th} head.

2.4 Spatiotemporal Information Filter

In addition to capturing the explicit spatiotemporal preference activated by the current search action, we further introduce the StIF module, which utilizes the search state representation to extract as much implicit spatiotemporal information as possible from the features. In this paper, StIF automatically identifies significant features with latent spatiotemporal effects from the user and context fields, depending on the feature construction of the specific recommendation system. To achieve this, we adopt the gating method GateNet [7], which incorporates a learnable feature gating unit to effectively capture high-order interactions. The core idea can be described as follows:

$$o_i = s_i \otimes \sum_j w_{i,j},\tag{5}$$

$$w_{i,j} = \Phi_w(s_i, z_j), \tag{6}$$

where s_i represents the spatiotemporal activation representation of the current search state, z_j denotes the features from the E_U and E_C fields, and Φ_w is an MLP layer and sigmoid function that generate feature importance weights through information interaction.

The final loss function is a summation of the CTR prediction loss and the spatiotemporal contrastive loss:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{CTR} + (1 - \alpha) \mathcal{L}_{CL},\tag{7}$$

where $\alpha \in [0, 1]$ is a hyperparameter used to balance the contributions of \mathcal{L}_{CTR} and \mathcal{L}_{CL} .

3 EXPERIMENTS

3.1 Experimental Setting

3.1.1 Datasets. The **Real-OFD Public Dataset**¹ is randomly collected from the retrieval logs of active users on the Ele.me OFD platform, comprising approximately 400,000 samples. It encompasses more than 30 features, including user behavior sequences spanning a maximum length of 50, recorded over a month. The **Industrial Dataset** is constructed using impression and click logs obtained from the Ele.me OFD platform. The training set consists of samples from the preceding 89 days, while the test set comprises data from the subsequent day. The user's click behavior within the past 45 days is considered as the historical behavior sequence.

3.1.2 Baselines. We compare six widely-used and industry-proven CTR models: Wide & Deep [1], DCN [13], DeepFM [6], DIN [16], DIEN [15], and AutoInt [10]. Additionally, to demonstrate the effectiveness of our proposed modules in CSPM, we apply the CSRL module to DIN and activate other features using StPE and StIF modules. Furthermore, to compare the learning ability for spatiotemporal features, we provide comparisons of three spatiotemporal CTR models: StEN [8], BASM [4], and TRISAN [9].

All models in this paper are trained using the same training strategy and features. We utilize AdagradDecay [5] as the optimizer for these datasets and apply exponential decay, starting with a learning rate of 0.01 for the Real-OFD Public Dataset and 0.001 for the Industrial Dataset. We set the feature embedding dimension to 32 and truncate the length of users' behaviors to 100.

For evaluating the model performance, we employ the widelyused metric AUC (Area Under ROC), which has been validated as a good measurement for CTR prediction [2]. AUC is insensitive to the classification threshold and the positive ratio, with a maximum value of 1 indicating best performance.

 $^{^{1}}https://tianchi.aliyun.com/dataset/dataDetail?dataId=120281$

Table 1: Model Performance (AUC) on the Industrial Dataset and Real-OFD Dataset

Model	Industrial	Real-OFD
Wide & Deep [1]	0.8570	0.7446
DCN [13]	0.8568	0.7449
DeepFM [6]	0.8576	0.7450
DIN [16]	0.8581	0.7452
DIEN[15]	0.8561	0.7442
AutoInt [10]	0.8574	0.7445
DIN+CSRL	0.8589	0.7459
DIN+StPE+StIF	0.8585	0.7453
DIN+CSRL+StPE	0.8585	0.7467
DIN+CSRL+StIF	0.8592	0.7466
DIN+CSRL+StPE+StIF	0.8601	0.7479
StEN[8]	0.8596	0.7469
BASM[4]	0.8586	0.7471
TRISAN[9]	0.8571	0.7459
CSPM	0.8605	0.7481

Performance Comparison 3.2

In this subsection, we compare CSPM with three types of methods. It is worth noting that a 0.1% improvement in AUC is considered quite significant on an online e-commerce platform.

As shown in Table 1, CSPM exhibits better performance compared to five mainstream CTR models. We observe that DIN, as one of the most commonly-used models in online systems, shows excellent performance on both datasets due to its ability to capture users' interests using the attention mechanism. However, the performance of other methods varies widely on both datasets. Besides, to investigate the generalization of our proposed modules, we conduct variation experiments by applying the CSRL, StPE, and StIF modules to DIN. As seen in Table 1, these modules contribute to improving the model's effectiveness. Specifically, the CSRL module shows a more significant improvement compared to the other two modules, indicating the importaness of the search state's spatiotemporal representation. It is noteworthy that DIN with all the modules achieves a performance similar to that of CSPM. Such result further demonstrates that our proposed spatiotemporal modules can be applied to existing models to capture users' spatiotemporal preferences. In the third group of experiments, we compare CSPM to three spatiotemporal-enhanced models. StEN and BASM outperform TRISAN, which lacks exploration of users' spatiotemporal behaviors. In summary, CSPM excels in capturing users' strong spatial and temporal preferences, thereby improving the performance of models.

Ablation Studies 3.3

To validate the effectiveness of our CSPM, we conducted ablation studies. Firstly, we examined the effect of our CSRL module by removing the contrastive loss and the DCN-v2 module, respectively. As shown in Table 2, the experimental results demonstrate the effectiveness of these two modules, particularly the contrastive loss, which facilitates the learning of valuable spatiotemporal representations. Although the StPE module outperforms the StIF module by

Table 2: Ablation Performance (AUC) the Industrial Datase
and Real-OFD Dataset.

Model	Industrial	Real-OFD
w/o CSRL (\mathcal{L}_{CL})	0.8557	0.7451
w/o CSRL (DCN-v2)	0.8569	0.7458
w/o CSRL	0.8556	0.7425
w/o StPE	0.8573	0.7461
w/o StIF	0.8596	0.7479
w/o StPE+StIF	0.8561	0.7439
w/o CSRL+StPE+StIF	0.8550	0.7412
CSPM	0.8605	0.7481

0.23% on the industrial dataset and 0.18% on the Real-OFD dataset, indicating that the effect of explicit spatiotemporal preferences is more noticeable than implicit spatiotemporal preferences, it is also evident that StIF, activated by CSRL, contributes to performance improvement, thereby confirming the effectiveness of capturing latent spatiotemporal effects.

Result from online A/B testing 3.4



Figure 2: The different improvements achieved by CSPM.

From January 20, 2023, to February 18, 2023, we conducted online A/B testing on the Ele.me OFD platform. The A/B tests were carried out using two buckets: one as the experimental group and the other as the baseline group. To maximize the confidence level of the A/B experiment, we randomly assigned users to the buckets. After one month of cumulative testing, each test bucket accumulated close to twenty million data points. Comparing CSPM with the online-serving base model, we observed an improvement in CTR by 0.88% and a 1.0% increase in the purchase rate of each impression. These improvements were verified through hypothesis testing. Fig 2 illustrates the different degrees of improvements achieved by CSPM under medicine and merchant brand two types of queries in all periods and delivery distances. CSPM has made significant advancements in the service requests that require spatiotemporal sensitivity, such as medical queries during office/night hours and within a 3 km distance. Currently, CSPM has been deployed online and serves the main traffic, resulting in a significant enhancement in user experience and a substantial growth in business revenue.

4 CONCLUSIONS

This paper presents CSPM, a solution for tackling the challenge of diverse user behavior and interests in different locations and times, which affects CTR prediction in OFD platforms. CSPM comprises CSRL, StPE, and StIF modules to model complex spatiotemporal preferences within multi-field features. CSRL transforms the current user search state into a spatiotemporal activation representation for subsequent spatiotemporal preferences. StPE, utilizing a multi-head attention mechanism and SAR, extracts diverse location and time preferences from the historical behavior sequence field. StIF incorporates a gating network to capture important features with latent spatiotemporal effects from user and context fields. These modules disentangle users' spatiotemporal preferences under different search states from multiple-field features. Extensive experiments on two large-scale industrial datasets demonstrate the superior performance of CSPM. A/B testing and online deployment confirm significant business improvement with CSPM.

REFERENCES

- Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems. 7–10.
- [2] 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.
- [3] Qiang Cui, Chenrui Zhang, Yafeng Zhang, Jinpeng Wang, and Mingchen Cai. 2021. St-pil: Spatial-temporal periodic interest learning for next point-of-interest recommendation. In Proceedings of the 30th ACM International conference on information & knowledge management. 2960–2964.
- [4] Boya Du, Shaochuan Lin, Jiong Gao, Xiyu Ji, Mengya Wang, Taotao Zhou, Hengxu He, Jia Jia, and Ning Hu. 2022. BASM: A Bottom-up Adaptive Spatiotemporal Model for Online Food Ordering Service. arXiv preprint arXiv:2211.12033 (2022).
- [5] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning* research 12, 7 (2011).
- [6] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. arXiv preprint arXiv:1703.04247 (2017).
- [7] Tongwen Huang, Qingyun She, Zhiqiang Wang, and Junlin Zhang. 2020. GateNet: gating-enhanced deep network for click-through rate prediction. arXiv preprint arXiv:2007.03519 (2020).
- [8] Shaochuan Lin, Yicong Yu, Xiyu Ji, Taotao Zhou, Hengxu He, Zisen Sang, Jia Jia, Guodong Cao, and Ning Hu. 2022. Spatiotemporal-Enhanced Network for Click-Through Rate Prediction in Location-based Services. arXiv preprint arXiv:2209.09427 (2022).
- [9] Yi Qi, Ke Hu, Bo Zhang, Jia Cheng, and Jun Lei. 2021. Trilateral spatiotemporal attention network for user behavior modeling in location-based search. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 3373–3377.
- [10] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via selfattentive neural networks. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 1161–1170.
- [11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).
- [12] Daheng Wang, Meng Jiang, Munira Syed, Oliver Conway, Vishal Juneja, Sriram Subramanian, and Nitesh V Chawla. 2020. Calendar graph neural networks for modeling time structures in spatiotemporal user behaviors. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2581–2589.
- [13] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD*'17. 1–7.
- [14] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-Scale Learning to Rank Systems. In *Proceedings of the Web Conference* 2021 (Ljubljana, Slovenia) (WWW '21). Association for Computing Machinery, New York, NY, USA, 1785–1797. https://doi.org/10.1145/3442381.3450078
- [15] 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.
- [16] 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.