

分类号: TP39

密 级: 公开

学校代码: 10363

学 号: 2220920103



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目 上下文增强的深度序列推荐
算法研究

论 文 作 者 高祥云

指 导 教 师 孔超 副教授

学 科 (专 业) 软件工程

研 究 方 向 领域软件工程

论文提交日期: 2025 年 5 月 9 日

分类号：TP39

单位代码：10363

密 级：公开

学 号：2220920103

题 目 上下文增强的深度序列推荐算法研究

英文并列

题 目 Research on Context Enhanced Deep Sequence Recommendation
Algorithm

学生姓名： 高祥云

指导教师： 孔超 副教授

专 业： 软件工程

研究方向： 领域软件工程

论文答辩日期： 2025 年 05 月 10 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：高祥云

日期：2025年5月9日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ☐，在_____年解密后适用本版权书。

本学位论文属于

不保密 ☒。

学位论文作者签名：高祥云

日期：2025 年 5 月 9 日

指导教师签名：孔庆华

日期：2025 年 5 月 9 日

上下文增强的深度序列推荐算法研究

摘 要

推荐系统广泛应用于个性化信息服务领域，能够有效缓解用户在海量信息中筛选内容的负担。其中，序列推荐利用用户历史交互行为建模动态偏好，相较于传统方法能更准确捕捉用户兴趣变化。然而，现有序列推荐方法仍面临数据稀疏和冷启动问题，导致推荐效果受限。此外，随着用户数据安全和隐私保护需求的提升，如何在提升推荐质量的同时保障数据安全，仍是该领域亟待解决的关键问题。

序列推荐通过建模用户的历史行为，以挖掘序列内部的项目转换模式，从而预测用户的未来兴趣。常见方法主要利用序列信息捕捉项目间的时序依赖关系，以提升推荐的连贯性和精准度。然而，这类方法大多局限于单一用户自身行为序列，忽略了细粒度时间信息对用户兴趣迁移的影响。在实际应用中，用户兴趣不仅受到自身历史行为的驱动，还可能受到社交好友偏好动态变化的影响；同样，项目特征的演变也可能受到其他用户行为序列中相关项目的潜在干预。此外，随着数据安全与用户隐私保护需求的提升，传统方法依赖大规模用户行为数据进行建模的方式面临数据可用性受限的问题，进一步制约了推荐效果。因此，如何充分融合跨领域数据和辅助信息，在提升推荐质

量的同时保障数据安全，仍是序列推荐面临的关键挑战。针对上述问题，本文进行了以下研究，具体工作如下：

(1) 针对现有序列推荐对时间信息利用不足的问题，提出了时间上下文增强的单域序列推荐模型。该模型是一种面向单域序列推荐场景，构建多层次时序建模框架，融合时间间隔感知自注意力机制与时间卷积神经网络，全面捕获用户序列中的时间依赖特征与全局交互模式。模型由三个部分组成：首先，通过位置编码与时间间隔矩阵增强项目序列的时间感知表示；其次，结合多头自注意力机制捕获跨时间步的全局交互依赖，进一步地，结合多层扩张因果卷积提取局部时间模式并实现多层特征融合；最后，根据注意力表示与时序特征的动态融合结果，计算下一项目的预测概率分布。

(2) 针对单域推荐无法全面捕捉用户跨平台兴趣迁移模式的问题，提出一种社交增强时间感知的跨域序列推荐模型。该模型由三个主要部分组成：首先，嵌入模块用于将时间戳和交互信息分别转换为嵌入向量；其次，图神经网络模块用于揭示序列间项目的协作特征，而序列注意力编码器则提取项目序列中的序列特征，并揭示不同上下文中时间信息的潜在依赖关系；最后，设计了一个具有社会感知的注意力网络，以考虑用户的社会邻居信息，并捕捉社会关系的动态模式，进而实现跨域序列推荐。

(3) 针对传统推荐系统依赖集中数据训练带来的隐私泄露风险，面向推荐系统提出一种支持隐私保护的端云协同训练框架。具

体分为三个部分：首先，将多种深度学习框架实现的云侧模型转换成通用 MNN 模型供端侧设备训练；其次，云侧将模型下发给端侧设备，端侧初始化后，获取本地数据进行训练并计算损失，再执行梯度反向传播；最后，端侧训练后的模型反馈给云侧，通过联邦学习框架进行模型聚合与更新，再根据不同需求，将云侧模型按需部署到端侧设备上，实现端云协同。

关键词：序列推荐，注意力机制，社交网络，图神经网络，隐私保护

RESEARCH ON CONTEXT ENHANCED DEEP SEQUENCE RECOMMENDATION ALGORITHM

ABSTRACT

Recommender systems have been widely adopted in personalized information services, effectively mitigating the burden on users when filtering vast amounts of information. Among these, sequential recommendation leverages users' historical interactions to model dynamic preferences, enabling a more precise capture of evolving user interests compared to traditional methods. However, existing sequential recommendation approaches still suffer from challenges such as data sparsity and the cold-start problem, which constrain recommendation performance. Furthermore, with the increasing emphasis on data security and user privacy protection, ensuring high quality recommendations while safeguarding data security remains a pressing issue in this field.

Sequential recommendation aims to predict users' future interests by modeling their historical behaviors to uncover the item transition patterns within sequences. Common approaches primarily leverage sequential information to capture the temporal dependencies between items, thereby enhancing the coherence and accuracy of recommendations. However, most of these methods are confined to a single user's own behavior sequence, overlooking the impact of fine-grained temporal information on user interest shifts. In practical applications, users' interests are not only

driven by their own historical behaviors but may also be influenced by the dynamic changes in their social friends' preferences. Similarly, the evolution of item characteristics may be subject to potential interventions from related items in other users' behavior sequences. Moreover, with the increasing demand for data security and user privacy protection, the traditional modeling approach that relies on large-scale user behavior data faces constraints in data availability, further limiting the recommendation performance. Therefore, how to effectively integrate cross-domain data and auxiliary information while improving recommendation quality and ensuring data security remains a key challenge in sequential recommendation. In response to the aforementioned issues, this paper conducts the following research, and the specific work is as follows:

(1) To address the insufficient utilization of temporal information in existing sequential recommendation methods, a temporal-aware convolutional sequential recommendation model is proposed. This model targets single-domain sequential recommendation scenarios, constructing a multi-level temporal modeling framework that integrates a time-interval-aware self-attention mechanism and temporal convolutional neural network to fully capture temporal dependencies and global interaction patterns within user sequences. The model consists of three main components: first, positional encoding and time interval matrices enhance the temporal awareness of item sequences; second, multi-head self-attention captures global interaction dependencies across time steps; third, multi-layer dilated causal convolutions extract local temporal patterns and achieve multi-level feature fusion. Finally, the next-item prediction probability distribution is calculated based on the dynamically fused attention representation and temporal features.

(2) To overcome the limitation of single-domain recommendation in capturing users' cross-platform interest transfer patterns, a social-augmented time-aware cross-domain sequential recommendation model is proposed. This model comprises three key components: first, an embedding module encodes timestamps and interaction information into embedding vectors; second, a graph neural network module uncovers collaborative features between items across sequences, while a sequence attention encoder extracts sequence features within item sequences and reveals potential dependencies of temporal information in different contexts. Finally, a social-aware attention network is designed to incorporate information from users' social neighbors and capture dynamic social relationship patterns, enabling cross-domain sequential recommendation.

(3) To mitigate the privacy leakage risks caused by centralized data training in traditional recommendation systems, a privacy-preserving edge-cloud collaborative training framework has been proposed for recommendation systems. This method consists of three parts: first, cloud-side models implemented using various deep learning frameworks are converted into a unified MNN model for training on edge devices; second, the cloud distributes the model to edge devices, which initialize the model, retrieve local data for training, compute the loss, and perform backpropagation; finally, the trained models from edge devices are fed back to the cloud, where a federated learning framework aggregates and updates the model. The updated model can be deployed back to edge devices as needed, enabling edge-cloud collaborative learning.

Gao Xiangyun (Software Engineering)

Supervised by Prof. Kong Chao and Wu Qilin

KEYWORDS: Sequence recommendation, Attention mechanism, Social network, Graph neural network, Privacy protection

目 录

摘 要.....	I
ABSTRACT.....	IV
第 1 章 绪论.....	1
1.1 研究背景及意义.....	1
1.2 研究内容.....	4
1.3 论文组织结构.....	6
第 2 章 背景知识及研究现状.....	8
2.1 序列推荐系统概述.....	8
2.2 国内外研究现状.....	9
2.2.1 序列推荐.....	9
2.2.2 社会化推荐.....	13
2.2.3 数据安全性与隐私保护.....	14
2.3 本章小结.....	15
第 3 章 时间上下文增强的单域序列推荐.....	16
3.1 研究背景.....	16
3.2 模型目标.....	17
3.3 模型结构.....	17
3.3.1 序列嵌入模块.....	18
3.3.2 时序特征构建模块.....	20
3.3.3 模型预测与训练.....	21
3.4 实验与分析.....	22
3.4.1 实验设置.....	22
3.4.2 性能比较.....	24
3.4.3 超参实验分析.....	25
3.5 本章小结.....	26
第 4 章 社交增强时间感知的跨域序列推荐.....	27
4.1 研究背景.....	27
4.2 模型目标.....	29
4.3 模型结构.....	30
4.3.1 嵌入初始化.....	30

4.3.2 图神经网络模块.....	31
4.3.3 注意力模块.....	32
4.3.4 模型预测与训练.....	33
4.4 实验与分析.....	34
4.4.1 实验设置.....	34
4.4.2 性能比较.....	36
4.4.3 消融实验分析.....	36
4.4.4 超参实验分析.....	38
4.5 本章小结.....	39
第 5 章 支持隐私保护的端云协同训练框架.....	40
5.1 研究背景.....	40
5.2 模型目标.....	41
5.3 模型结构.....	42
5.3.1 云侧模型准备.....	42
5.3.2 端侧模型部署与训练.....	43
5.3.3 端云服务.....	43
5.3.4 端云协同训练算法实现.....	45
5.4 实验与分析.....	46
5.4.1 框架功耗测试.....	46
5.4.2 DSSM 与 Deep&Wide 模型端云协同训练.....	47
5.6 本章小结.....	49
第 6 章 总结与展望.....	50
6.1 工作总结.....	50
6.2 未来展望.....	51
参考文献.....	52
攻读学位期间发表的学术论文目录.....	60
致 谢.....	61

第 1 章 绪论

1.1 研究背景及意义

在数字时代的快速变革中，信息量的剧增使得用户在网络世界中获取有效信息。据中国互联网络信息中心（CNNIC）第 55 次《中国互联网络发展状况统计报告》显示，截至 2024 年 12 月，我国网民规模近 11.08 亿人，互联网普及率达 78.0%^[1]。现如今，互联网上不同领域涌现出大量的商品、新闻、视频等多样性内容，构成了一股庞大的信息洪流，在这种情况下，用户很难提取到需要的信息，从而产生了信息过载问题^[2-5]。这种信息过载的问题使得用户不仅需要更智能、个性化的信息过滤工具，同时也需要更具洞察力和适应性的信息获取策略，以便更好地在这个信息爆炸的时代中高效地获取有价值的内容。

为了应对信息过载的挑战，推荐系统作为一种精准且个性化的信息推荐工具，得到了广泛的关注和重视。通过分析用户的行为、兴趣和偏好，推荐系统能够帮助用户从海量的信息中筛选出相关内容^[6,7]，提供量身定制的推荐服务，有效缓解信息过载并提高信息获取效率。传统搜索引擎通常依赖关键词匹配来检索与用户查询相关的网页信息，而推荐系统则无需用户直接表达需求，通过挖掘用户的历史行为模式构建个性化模型，从而提供更符合用户兴趣的推荐内容。这种主动且个性化的推荐方式帮助用户更高效地筛选信息，缓解信息过载的问题，同时提升用户获取信息的效率^[8]。在众多推荐系统的细分领域中，序列推荐系统相比传统推荐方法，更加注重分析和建模用户的持续交互行为、用户与项目的互动、项目的流行度以及用户偏好的时间演化。通过深入挖掘序列中所蕴含的依赖关系和动态相关性，序列推荐系统能够更精准、个性化和动态地生成推荐结果^[9]在实际场景中，用户的交互行为序列是相对容易获取的，因此序列推荐的研究更符合当前的应用场景。此外，将用户的偏好分为长期和短期考虑具有重要的现实意义，因为用户的短期偏好和长期偏好往往具有一定的联系或存在较大的差异。因此，序列推荐系统在分析用户偏好方面具有很大的潜力，有望为现实生活中的推荐领

域带来更多的创新和突破。

虽然个性化序列推荐在一定程度上缓解了信息过载的问题,但传统的序列推荐方法仍然面临以下问题:(1)数据稀疏:推荐系统中普遍存在数据稀疏问题,是指用户与项目的交互数据(如评分、点击、购买等)极度匮乏,导致系统难以从有限信息中准确建模用户偏好或项目特征的现象;(2)冷启动:冷启动是推荐系统因缺乏历史交互数据(如用户行为、项目特征或系统初始数据)而难以生成有效个性化推荐的初始阶段问题;(3)数据安全性与隐私保护问题:随着对用户数据的日益依赖,推荐系统也面临着用户数据安全性与隐私保护的挑战。推荐系统在处理用户数据时,应严格保护用户隐私,避免信息外泄或被滥用。研究者往往通过引入多种辅助信息缓解数据稀疏与冷启动问题,如项目内容特征、社交网络关系、知识图谱等。此外,针对数据安全性与隐私保护的挑战,研究者采用了诸如数据加密、差分隐私、联邦学习等隐私保护技术,以确保用户数据的安全性和隐私得到充分保护。

在这些解决方案中,社交网络信息逐渐成为一种重要的辅助信息来源。随着社交媒体的普及,用户之间的关系变得更加丰富。这些社交关系不仅能够揭示用户之间的潜在联系,还可以作为推荐系统建模用户兴趣和行为的重要依据。通过利用社交关系,推荐系统能够缓解数据稀疏和冷启动问题,并提升个性化推荐的效果。然而,社交关系和用户兴趣并非静态不变,而是随着时间的推移不断演化。传统的序列推荐方法往往忽略了社交关系的动态性,难以适应用户兴趣的持续演化,从而导致推荐结果滞后,无法满足用户的个性化需求。

除了社交数据以外,用户的行为轨迹还广泛分布在电商平台、搜索引擎、资讯平台等多个应用场景中,形成了多源异构数据。用户的社交互动、购物记录和搜索历史等行为数据,往往来源于不同领域,并呈现出高度的异构性和碎片化。现有的序列推荐研究主要聚焦于利用单个领域(或称为单一数据源)的数据进行推荐,这种方法在捕捉特定领域内用户行为模式方面取得了一定的成效。然而,仅依赖单个域的数据进行推荐存在明显的局限性。首先,单个域的数据可能无法全面反映用户的真实兴趣和需求。用户的兴趣往往跨越多个领域,例如,一个用户可能同时对服饰和电子产品感兴趣,但这些兴趣可能分散在社交网络和电商平

台等不同领域的的数据中。其次，单个域的数据可能存在稀疏性问题，特别是在用户历史行为记录较少的情况下，难以准确构建用户画像和进行精准推荐。此外，随着推荐系统的发展，不得不面对用户隐私泄露的严重威胁。尽管有关序列推荐系统的研究取得了显著的进展，但对用户隐私的保护仍是一个需要重视的问题。

为缓解这些问题造成的影响，本文提出了一种创新的序列推荐框架，旨在突破现有方法在单一用户行为序列建模中的局限，融合细粒度的时间上下文信息和社交邻居数据，提升推荐系统的准确性与冷启动性能。该框架涵盖了从时间上下文信息感知建模、社交关系增强与跨域序列建模到隐私保护的端云协同训练框架的全过程，旨在通过融合多层次时间依赖特征、社交动态演化和隐私安全保障，全面提升推荐系统的性能与用户体验。

尽管现有序列推荐方法通过时序模式挖掘提升了短期行为预测的精度，但在稀疏交互数据环境下，其泛化能力仍然受限，尤其是在冷启动对象的推荐置信度方面存在显著下降。此外，传统静态建模假设难以适应现实场景中用户兴趣的变化、社交网络结构的动态调整以及项目特征的时效性衰减，导致长期推荐效果不稳定。同时，隐私保护在推荐系统中同样具有重要意义，需要在个性化推荐的同时保障用户数据安全。本研究存在以下四点挑战：

（1）数据稀疏问题：数据稀疏问题是推荐系统中常见的挑战，即用户与项目之间的交互数据量较少，使得推荐系统难以有效学习用户偏好。例如在一个电商平台上，商品种类繁多，但大多数用户只购买或浏览其中的少量商品，这种交互信息的缺乏，使得系统在面对大量未浏览或未购买的商品时，难以准确预测用户的兴趣。由于用于训练推荐算法的数据量严重不足，系统很容易陷入过拟合的困境，且稀疏数据会导致推荐的准确性下降。

（2）冷启动问题：在推荐系统中，历史交互数据的缺乏，特别是在新用户注册或新项目添加时，导致系统难以收集足够的信息进行精准建模。这种数据匮乏的情况使得对新用户的兴趣偏好或对新项目的潜在受欢迎程度进行有效预测变得非常困难，从而影响了推荐的准确性和效果。

（3）社交关系的多样性和影响力差异：在实际应用中，用户之间的社交关系往往呈现多样性，并且这些关系的影响力存在差异。不同用户的社交网络对其

兴趣迁移的影响程度不同，这导致了社交推荐系统中的信息传递并非均匀。某些社交关系可能在推荐系统中发挥更大作用，而其他关系则可能较为薄弱或不相关。因此，如何精准捕捉和建模社交网络中有影响力的关系，以及如何利用这些关系对用户的兴趣变化进行有效建模，是一个亟待解决的挑战。

(4) 数据安全与隐私保护问题：推荐系统在各个领域的广泛应用使得用户隐私安全问题日益严重。这些系统依赖于大量的用户历史行为数据和交互记录来训练模型，这些数据可能包含敏感信息，如用户的兴趣偏好、社交关系、地理位置等。一旦用户数据被滥用或泄露，不仅可能侵犯用户隐私，还可能导致恶意推送或信息操控等问题。因此，在推荐系统中引入隐私保护机制，确保用户数据在传输、存储和处理过程中的安全性，已成为学术界和工业界广泛关注的研究方向。

1.2 研究内容

本文将社交信息和时间间隔数据作为辅助信息，同时对推荐系统中的隐私保护问题进行研究，挖掘和表征了用户的偏好，并将其融入序列推荐模型中以提升推荐效果。



图 1-1 研究框架

Fig.1-1 The research framework

本文的研究围绕上下文辅助信息和数据安全与隐私保护的序列推荐算法展开,重点探索以下三个递进式研究内容:分别是时间上下文增强的单域序列推荐,以及社交增强时间感知的跨域序列推荐和支持隐私保护的端云协同训练框架。研究框架图如图 1-1 所示,主要包含如下内容:

(1) 时间上下文增强的单域序列推荐 (temporal-aware convolutional sequential recommendation, TCSRec)。近年来,序列推荐的研究主要聚焦于长短期兴趣建模,将同一用户的多个交互按时间顺序排列进行建模,相比传统方法,能够更好地学习用户的动态兴趣。然而,现有方法往往对时间上下文信息的建模仅局限于位置编码或简单时间差计算,未能充分挖掘序列中隐含的时间依赖关系。本文提出一种多层次时序建模框架的解决方案,模型由三个部分组成:首先,通过位置编码与时间间隔矩阵增强项目序列的时间感知表示;随后,结合多头自注意力机制捕获跨时间步的全局交互依赖,进一步地,结合多层扩张因果卷积提取局部时间模式并实现多层特征融合。最终,根据注意力表示与时序特征的动态融合结果,计算下一项目的预测概率分布。

(2) 社交增强时间感知的跨域序列推荐 (social-augmented time-aware cross-domain sequential recommendation, STRec)。现有的序列推荐方法主要聚焦于利用单一领域数据进行推荐,但为了更好地理解用户的兴趣与行为,并进一步提升推荐系统的性能,跨域推荐方法已经成为推荐系统领域的一个重要研究方向。基于此,本文提出了一种多源异构数据的跨域序列推荐算法,能够自动捕捉多源异构信息的特征,进而实现这些信息的有效提取与融合,以缓解数据稀疏和冷启动问题。此外,通过引入社交上下文信息作为辅助特征来建模用户兴趣,并将用户的社交上下文信息应用于不同领域的推荐任务中,以增强推荐的多样性和准确性。为此,本文进一步提出了一个社交增强时间感知的跨域序列推荐模型。该模型由三个核心部分组成:首先,通过嵌入模块将时间戳和交互信息分别转换为嵌入向量,便于处理时间依赖特性;其次,图神经网络模块用来揭示不同领域间项目的协作特征,而序列注意力编码器则从项目序列中提取序列特征,并分析不同上下文中时间信息的潜在依赖关系。最后,设计了一个具有社会感知的注意力网络,通过考虑用户的社会邻居信息,捕捉社交关系的动态演化,从而提升跨域推荐的

效果。

(3) 支持隐私保护的端云协同训练框架 (federated machine learning and mobile neural network, FedMNN)。在个性化推荐日益普及的同时, 用户隐私泄露的威胁变得越来越严重。尽管有关序列推荐系统的研究取得了显著的进展, 但对用户隐私的保护仍是一个需要重视的问题。因此提出了一种能够面向推荐系统的支持隐私保护的端云协同训练框架, 能够在不泄露用户隐私的前提下传输数据信息。联邦学习作为一种分布式机器学习技术, 能够在不需要交换本地个体或样本数据的前提下进行数据传输, 有效保护了用户的数据隐私。基于联邦学习, 本文提出了一个可用于序列推荐任务的端云协同训练框架, 该框架结合了联邦机器学习与移动神经网络, 将多种深度学习框架实现的云侧模型以转换成通用模型供端侧设备训练, 云侧将初始化的模型下发给端侧进行训练, 端侧设备使用本地的用户-项目交互序列信息进行序列推荐模型训练。每轮训练结束后, 端侧将训练好的模型参数上传至云侧, 通过联邦学习聚合框架进行模型参数聚合与更新, 确保了模型能够整合端侧设备的个性化信息, 形成更全面的推荐模型。这个框架通过端侧和云侧协同训练, 在保证模型性能的前提下, 避免个人隐私数据泄露, 并有效解决数据孤岛问题。

1.3 论文组织结构

本文的研究内容和组织结构如下:

第一章围绕上下文增强的深度序列推荐算法, 阐述研究背景及其意义, 分析当前序列推荐系统所面临的关键问题, 并概述研究框架和主要贡献。最后, 对本文的组织结构进行了总结。

第二章介绍了序列推荐的相关理论基础, 并梳理国内外在本研究涉及的核心技术领域的研究进展, 对序列推荐算法、社会化推荐算法及数据安全与隐私保护的国内外研究做出了介绍。

第三章研究时间上下文增强的单域序列推荐算法, 提出基于时间上下文增强的单域序列推荐方法。在此章节中, 首先对算法的整体框架进行了详细阐述, 接

着逐一分析了算法中各个模块的设计与实现。通过一系列真实世界数据集上的对比实验，展示了该模型在推荐效果上的显著提升。

第四章研究社交增强时间感知的跨域序列推荐算法，提出社交增强时间感知的跨域序列推荐方法。本章首先介绍了该算法的整体框架，随后对各个模块的设计与实现进行了详细分析。最后，通过在两个数据集上的对比实验，验证了该算法在推荐效果上的有效性。

第五章研究支持隐私保护的端云协同训练框架，面向推荐系统提出了一种结合联邦机器学习与移动神经网络的端云协同训练框架，首先介绍了该训练框架的整体流程与系统架构，然后分别对各个模块的功能和工作原理进行介绍，通过两个数据集上的对比实验证明了该训练框架的有效性。最后展示了该训练框架的原型系统。

第六章再次对本研究的内容进行了总结，表明了本研究的优势和不足之处，并对下一步的优化方案进行了初步探究，最后对本研究的未来发展方向进行了展望。

第2章 背景知识及研究现状

本章主要介绍了序列推荐系统相关的背景知识以及研究现状。对序列推荐做出简单概述，介绍了序列推荐的背景、所用相关技术以及应用场景。然后，介绍了与本研究相关的国内外研究现状。

2.1 序列推荐系统概述

随着互联网的快速发展，信息爆炸现象日益显著，用户在网络世界中获取信息的难度逐渐加大，导致“信息过载”问题愈发严重。推荐系统的出现，为缓解信息过载问题提供了一种有效的解决方案。在传统的推荐系统中，研究人员通常通过协同过滤、基于内容的推荐或矩阵分解等方法，挖掘用户的兴趣偏好并进行推荐。然而，这些方法主要依赖于用户的静态兴趣或长期历史行为，忽视了用户兴趣随时间变化的动态特征，导致推荐结果滞后，难以满足用户实时变化的需求。因此，如何动态捕捉用户的短期兴趣变化，并基于历史行为进行个性化推荐，成为推荐系统研究中亟待解决的问题。在此背景下，序列推荐应运而生，成为当前推荐系统研究的重要方向之一。

序列推荐的核心在于将用户历史交互行为视为时间序列，通过建模行为之间的转移规律，预测其下一个可能感兴趣的项目。与传统推荐关注“用户长期偏好”不同，它更强调“在当前行为轨迹下用户的即时需求”。这种视角转变使推荐系统具备了动态推理能力：当用户在购物车中添加跑步鞋后，系统会根据常见行为链“鞋类→运动袜→护膝”推荐关联商品。这种技术不仅需要理解单个项目的属性，更要挖掘项目在序列中的上下文关系。例如，在新闻推荐中，用户早晨浏览财经新闻与晚间阅读娱乐资讯可能反映出工作日与休息日的行为差异，序列模型需识别此类周期性模式，动态调整推送策略。因此，序列推荐本质上是通过对用户行为轨迹的“解码”，实现从静态偏好匹配到动态场景适应的跨越。目前，序列推荐已广泛应用于电商、短视频、音乐、新闻等领域，通过分析用户的浏览、点击、购买或播放记录，动态调整推荐内容，从而提升用户体验和平台收益。

序列推荐的关键技术融合了多种深度学习和图神经网络方法，具体实现中，主要涉及用户行为序列建模、全局依赖关系学习和多源信息融合。首先，通过对用户历史行为的建模，可以有效捕捉兴趣的动态变化，特别是利用循环神经网络（recurrent neural network, RNN）、自注意力机制等方法，提取时间序列中的短期和长期依赖特征，提升推荐的时效性和精准度。其次，图神经网络（graph neural network, GNN）的引入，使得系统能够从行为序列中挖掘项目之间的高阶关联，通过图卷积网络（graph convolutional networks, GCN）或图注意力网络（graph attention networks, GAT）等技术，捕捉复杂的行为转移模式，增强对用户偏好的刻画。此外，序列推荐还可以利用多种辅助信息来提升效果，如跨域行为、社交关系、文本描述等辅助信息，能够进一步丰富用户兴趣的表达，提高推荐的泛化能力。结合这些技术，序列推荐能够更精准地理解用户的行为轨迹，在动态场景中提供高效、个性化的推荐服务。

2.2 国内外研究现状

随着信息技术的快速发展和用户行为数据的持续积累，序列推荐作为推荐系统的重要研究方向之一，凭借对用户兴趣动态演化的有效建模，受到广泛关注。为进一步提升推荐性能，研究逐步扩展到跨域序列推荐、结合社交关系增强序列推荐以及数据安全和隐私保护等领域，以应对数据稀疏、冷启动和隐私泄露等挑战。本文围绕序列推荐展开研究，国内外相关工作主要从序列推荐算法、社会化推荐算法和数据安全与隐私保护三个方面进行总结与分析。

2.2.1 序列推荐

在推荐系统中，序列推荐是一种针对用户连续交互行为进行建模和分析的方法。这种推荐方法可以根据用户的历史行为序列来建模用户的动态偏好，从而能够比传统的推荐方法更加及时地捕捉用户的兴趣变化，因此，序列推荐的研究已经成为一个热门的话题^[10]。

（1）基于传统方法的序列推荐算法

传统序列推荐主要分为基于马尔可夫链^[11]（markov chain, MC）方法和基于矩阵分解^[12]方法。假设用户未来交互的项目与前一个或几个项目有关，在序列推荐任务中，马尔可夫链方法通过分析用户的历史行为序列，构建一个转移矩阵，并利用该矩阵中的转移概率来预测用户的下一个可能的交互行为。Shani 等人^[13]首次提出马尔可夫链方法，将推荐任务视为序列决策问题，以考虑推荐对用户长期行为的影响来解决序列推荐问题。Zhang 等人^[14]提出了一种结合马尔可夫模型和基于内容搜索技术的推荐方法，其中通过混合一阶和二阶马尔可夫模型对用户点击流进行建模，以缓解马尔可夫模型覆盖率低的问题，从而提升推荐效果。He 等人^[15]使用高阶马尔可夫链来对用户行为序列进行建模，实现了个性化的马尔可夫转换。FPMC^[16]将矩阵分解和马尔可夫模型相结合，实现了个性化序列推荐。FISM^[17]模型将项目-项目相似矩阵分解成两个低秩矩阵，有效地解决了模型在数据稀疏情况下性能下降问题。

马尔可夫链推荐方法虽然擅长建模用户的短期偏好，但在捕捉序列中的长期依赖关系方面存在局限性。类似地，矩阵分解技术应用于序列推荐时也存在不足，例如其通常仅关注低阶交互模式，难以充分挖掘高阶关系和时序特征，影响推荐效果的全面性。

（2）基于深度学习的序列推荐算法

随着深度学习技术不断发展，研究人员将循环神经网络、卷积神经网络（convolutional neural networks, CNN）、图神经网络^[18-20]等算法应用在序列推荐领域并取得了较好成果。例如，Hidasi 等人^[21]提出了 GRU4Rec 模型，首次将门控循环单元（Gated Recurrent Unit, GRU）引入到序列推荐中建模序列项目之间的关系。随后在其原有模型上通过不同的损失函数和采样策略提出了 GRU4Rec+^[22]模型。CARE^[23]模型使用 CNN 从序列中提取序列模式和用户偏好并用于评分预测，使用卷积注意力来学习序列模式。SRGNN^[20]、MTD^[24]和 GCE-GNN^[25]将项目序列转换为图，并使用图神经网络来捕获复杂的项目-项目转换。基于 RNN 或 CNN 的方法不能捕获在会话内和跨会话中用户兴趣的层次性质。此外，CNN 虽然具有并行计算能力，但难以有效刻画序列行为的层次特征，并

且对长序列建模时存在较高的内存消耗。针对上述问题, HierTCN^[26]利用 RNN 结合时间卷积网络模型, 预测下一次交互行为, 提高内存的使用效率及训练速度。王鸿伟等^[27]提出循环记忆网络 (recurrent memory networks, RMN), 分别构建记忆读取和写入单元, 捕获用户短期兴趣波动和长期兴趣转移。冯永等^[28]提出长短兴趣多神经网络混合动态推荐模型, 分别使用 RNN 和前馈神经网络, 对用户短期兴趣和长期偏好进行建模, 通过模型融合进行动态推荐。基于注意力的模型^[29-31]被应用于序列推荐任务中, 取得了出色的性能。SASRec^[32]应用 NLP 任务中的 Transformer 层来适应序列推荐。BERT4Rec^[33]则进一步考虑了双向 Transformer 层。

尽管上述模型能够较好地捕捉序列模式的动态变化, 但随着时间的推移, 用户的兴趣和行为模式会发生不断演化。传统的序列推荐模型往往忽略了这种长期变化, 仅关注用户的历史行为序列, 而缺乏对时间动态的有效建模。尤其是模型通常无法区分短期行为与长期偏好的变化, 导致在预测用户未来行为时无法准确反映出他们兴趣的波动。此外, 大部分现有方法在处理时间序列时未能充分考虑时间间隔不同的影响, 限制了模型对时间敏感特征的捕捉能力, 进而影响推荐结果的时效性和精准度。时间信息作为影响用户行为的重要因素, 也受到广泛关注。Wu 等^[34]提出 RRN, 将绝对时间戳作为辅助信息, 通过长短期记忆网络 (long short-term memory, LSTM) 构建自回归模型, 动态捕获用户和项目表示。Liu 等^[35]设计短期注意/记忆优先模型, 从会话上下文中获取用户长期记忆, 同时结合最后一次点击的短期兴趣, 实现更精细的偏好刻画。但上述模型未充分考虑时间信息的层次性和多尺度特征。为此, 袁涛等^[36]构建尺度维卷积网络, 从不同时间尺度建模用户兴趣演变。Li 等^[37]进一步引入时间间隔信息, 结合自注意力机制共同学习项目间的内容关系、位置关系与时间关系。Ji 等^[38]提出多跳时间感知注意力记忆网络, 设计时间感知门控循环单元, 分别捕获长期和短期兴趣, 并动态学习时间感知的短期意图。考虑到之前的方法仅考虑单一类型的行为, 如何有效地利用动态多行为知识来提高推荐准确性, 仍然是一个重大挑战。Xia^[39]等人提出时态图 Transformer 框架探索不同类型行为之间不断变化的相关性, 共同捕获动态的短期和长期用户项目交互模式。MBHT^[40]通过低秩自注意力, 可以从细粒

度和粗粒度级别联合编码行为感知序列模式,捕获短期和长期跨类型行为依赖性。MB-ATMK^[41]通过整合用户多行为交互、时序感知编码、属性权重增强的元知识图神经网络以及偏好迁移机制,有效捕获用户的动态偏好、细粒度属性信息和跨行为依赖。PBAT^[42]通过生成动态行为模式和融合行为-时间影响的协作提取器,有效建模多行为序列中的个性化模式与复杂依赖关系,从而提升多行为序列推荐的性能和可解释性。然而,上述方法没有考虑到序列中重要的上下文信息,即不同用户所处环境、实时需求等以及个人偏好等因素之间存在差异,且仍存在数据稀疏冷启动等问题,这可能导致推荐的准确性受到影响。

(3) 基于跨域信息的序列推荐算法

推荐系统通过引入各种辅助信息,如社交关系、项目特征和用户行为数据,有效缓解了“数据稀疏”和“冷启动”等问题。不同领域之间存在多种异构数据,跨域推荐的出现能够有效利用这些异构数据进行推荐,很好的缓解了“数据稀疏”问题。典型的跨域推荐模型是从单域推荐模型扩展而来的,例如 M³Rec^[43]。这些方法利用其他辅助域的交互信息来满足用户行为,从而在目标域中做出更好的推荐。随着迁移学习的出现激发了许多研究者基于迁移学习研究跨域推荐,例如 CCTL^[44]和 BiTGCF^[45]。这些方法首先利用两个基本神经网络分别对源域和目标域中的用户-项目交互进行建模,然后设计不同的传输模块来融合它们。跨域序列推荐能够利用不同域的信息进行序列推荐,典型的工作是 π -Net^[46]和 PS-JNet^[47],这些方法设计了复杂的门控机制来传输单域信息,利用并行的 RNN 组成并行信息共享网络,通过共享账户作为知识迁移桥梁,在共享账户过滤单元的帮助下提取特定用户的表示以识别不同的用户行为。Zhuang^[48]等人提出了一种跨域新颖搜寻算法,以更好地建模个人跨不同领域知识转移的倾向。一些工作在不同领域项目之间构建交互图(或知识图)来指导跨领域的信息传输,例如 DA-GCN^[49]和 MIFN^[50]。然而,上述跨域序列推荐方法大多是特定的多用户共享账号的跨域推荐问题,即假设多个用户共享一个账户并在不同领域中进行交互,但在实际应用中,用户在不同领域的行为往往是独立的,且每个领域的用户偏好和行为模式也可能有所不同。因此,这类方法难以适应更广泛的跨域推荐需求。此外,它们忽略了序列之间的项目关系,这可能导致在推荐过程中丢失了有关不同领域之间用

户偏好的重要信息。

2.2.2 社会化推荐

社会化推荐旨在利用社交网络数据中的社交关系信息以及用户-项目交互信息，社交关系所隐含的用户相关性为用户建模提供了额外的辅助信息，能够获得更准确的用户偏好，为每个用户生成个性化推荐，从而缓解冷启动和数据稀疏问题^[51]。早期的解决方案基于矩阵分解模型^[52]。随着深度学习的进步，解决方案逐步发展为神经社交推荐^[53]。为了更准确的建模用户，ARMF^[54]首次提出伪相似朋友的概念，通过构建用户-项目异构网络，计算用户间的相似度，挖掘用户间的相似关系，并引入注意力机制分析用户对项目属性特征的关注度。SoRec^[55]是一种基于概率矩阵分解的社交推荐方法，将用户社交网络结构与用户-项目评分矩阵通过共享的用户潜在特征空间相结合，实现朋友偏好信息的推荐。TrustMF^[56]是一种融合用户评分数据与社交信任网络的社交协同过滤方法，通过矩阵分解技术将用户映射到低维特征空间，从而捕捉用户之间的相互影响并改善冷启动用户的推荐效果。SocialM^[57]将信任传播机制引入推荐模型中，以利用社交网络中的用户信任关系，该模型的独特之处在于强调了社交网络中用户之间的兴趣相似性。SocialCDL^[58]提出运用多层神经网络和矩阵分解技术处理社交信息以进行协同过滤。特别地，SocialCDL 设计一种改进的深度自编码器模型 SSDAE^[59]以解决社交网络的数据稀疏及不平衡问题。Wan^[60]等人提出一种基于深度学习的新型信任感知方法，该方法引入了两种深度矩阵分解技术，以提升初始化精度，并将偏好相似性和派生关系建模为社会信任因素，进一步融入矩阵分解框架，以增强推荐性能。EAGCN^[61]提出了一种高效的自适应图卷积网络框架，通过引入空间自适应图卷积模块，联合探索用户兴趣传播和社交影响。尽管这些方法的结果颇具前景，但许多现有的社交推荐模型忽略了用户活动的序列动态，可能导致表现不尽如人意。根据现有研究，利用社交关系增强序列推荐性能的研究非常有限。一个近期的例子是 DGRec^[62]，该方法结合了 LSTM 单元来表示动态序列行为，并采用基于注意力的 GNN 来考虑静态社交影响。RP-SANRec^[63]使用 RNN 模拟用户偏好的变化，并结合注意力机制来检测社交影响。

2.2.3 数据安全和隐私保护

推荐系统可以在大规模数据中挖掘有价值的信息,缓解信息过载问题。推荐系统通常需要收集用户和项目的属性信息以及他们之间的交互记录,以便了解用户的偏好,从而提供精确的推荐。然而,这些用户属性和交互记录包含了大量的个人隐私信息,真实场景中往往存在隐私泄露的风险,因此,数据安全和隐私保护得到研究人员的重视。Hu 等人^[64]从添加随机扰动角度提出了一种多层次的隐私保护组合模型,根据高斯和均匀混合扰动动态划分多个扰动级别,实现矩阵分解模型的隐私保护。而基于差分隐私的技术^[65]主要思想是通过添加随机噪音扰乱原有数据,使得攻击者无法区分或辨别目标节点是否在数据集中,从而实现隐私保护。Zhang 等人^[66]针对加权社交网络的隐私泄露问题,提出了一种基于差分隐私的隐私保护推荐算法。传统的数据加密方法在一定程度上保护了用户的隐私,但也一定程度上阻碍了数据的流通,出现了“数据孤岛”^[67]问题。联邦学习是一种分布式机器学习技术,在保证各参与方数据隐私不被泄露的情况下,通过共享加密的中间计算结果,实现协同模型训练。这种联合训练方法保证模型性能的前提下,避免个人隐私数据泄露,并有效解决数据孤岛问题。谷歌于 2019 年推出了移动设备端联邦学习框架^[68],推动联邦学习在移动智能设备上的实验。微众银行提出了开源联邦学习框架 FATA^[69],实现多方安全模式下的机器学习、深度学习和迁移学习。美国南加州大学、麻省理工学院和斯坦福等高校和公司联合发布的联邦学习开源框架 FedML^[70],FedML 支持线性模型(如逻辑回归)和神经网络(如 CNN 和 RNN)根据训练设备的异构性,采用了消息队列遥测传输协议,实现云服务器与终端设备之间的通信。Flower^[71]是一款由英国牛津大学于 2020 年发布的联邦学习框架,它可以模拟真实场景下的大规模联邦训练,充分考虑计算资源、内存空间和通信资源,高效利用移动设备和无线客户端下异构资源^[72]。Wang 等人^[73]提出了一个具有联邦学习和隐私保护功能的跨领域推荐 POI 框架,采用联邦学习框架对用户本地历史数据进行分析,加密潜在特征分布进行知识迁移,以保护用户的隐私信息。研究发现,联邦学习框架在梯度传输时仍然可能会有隐私泄露的风险,甚至会泄露原始数据^[74],由此研究人员结合了梯度安全聚合

技术，如差分隐私^[75]、同态加密^[76]等。Truex 等人^[77]提出了一种满足差分隐私保护的联邦学习方法，在以高精度生成机器学习模型的同时能有效防止个体隐私泄露。然而，差分隐私中的模型性能与隐私保护程度之间难以平衡是个巨大的挑战，同态加密的计算及通信代价比较高，难以应用到真实环境中资源受限的地方。

2.3 本章小结

本章全面介绍了序列推荐系统相关的背景知识以及本文涉及到的相关研究现状，首先概述了序列推荐的核心概念及其关键技术。随后，系统梳理了国内外研究现状，重点探讨了基于传统方法的序列推荐、基于深度学习的序列推荐以及结合跨域信息的序列推荐，并分析了各类方法的优势与局限性。此外，还介绍了与序列推荐密切相关的社会化推荐，讨论其在利用社交关系增强推荐效果方面的研究进展，为后续结合社会关系优化推荐提供参考。同时，针对推荐系统中日益突出的数据安全与隐私保护问题，分析了现有的解决方案及发展趋势。通过本章的综述，为后续章节的研究奠定了理论基础，并提供了重要的研究背景与技术支撑。

第3章 时间上下文增强的单域序列推荐

考虑到序列推荐存在仅考虑用户行为的顺序,忽略用户行为之间的时间间隔信息的问题,本章提出了一种时间上下文增强的单域序列推荐方法 TCSRec,旨在有效捕捉用户行为的时间动态,提升推荐的精准度和时效性。3.1 节分析了现有序列推荐方法的局限性,并针对性地提出改进方案。3.2 节对序列推荐问题进行了形式化定义。3.3 节首先概述了 TCSRec 的整体框架,由三个模块组成,即嵌入表示模块、时序特征构建模块和模型预测与训练,随后对各模块的功能与工作原理进行详细解析。3.4 节在真实数据集上的实验展示出本章所提出的模型与现有对比方法的性能,并对结果进行分析。

3.1 研究背景

传统的序列推荐模型将用户交互按时间顺序排序,并通过挖掘序列模式来预测用户的下一次交互行为。典型解决方案包括基于马尔可夫链的方法^[78],其中 L 阶马尔可夫链根据前 L 个动作进行推荐。这类方法在高稀疏性场景下表现较优,得益于其简化假设,但在更复杂的环境中可能难以精准建模动态变化模式。循环神经网络^[79]也被应用于此场景,虽能长期记忆用户偏好,但需要大量数据(尤其是密集数据)才能超越简单基线。为克服马尔可夫链和 RNN 的局限性,受机器翻译中 Transformer^[80]的启发,研究者将自注意力机制引入序列推荐^[81,82],此类模型显著优于传统的 MC/CNN/RNN 方法。

现有的序列推荐模型往往忽略时间戳信息,仅关注项目的顺序,这隐含地假设所有相邻项目的时间间隔相同。这类方法主要依赖历史项目的位置和身份来预测用户的下一次交互。然而,在现实情况下,时间间隔较短的交互通常对后续行为的影响更为显著。例如,假设两名用户都观看了相同的电影序列,但一名用户在短时间内连续观看,而另一名用户则隔了数周才完成观看,他们对下一部电影的偏好可能有所不同。然而,传统序列推荐方法仅基于顺序信息建模,会将这两种情况视为相同,可能导致推荐结果缺乏个性化和时效性。

针对以上问题，本章提出一种面向时间上下文增强的单域序列推荐方法 TCSRec，该方法结合时间间隔感知自注意力机制与时间卷积神经网络，构建多层次时序建模框架。该模型由三个部分组成：序列嵌入模块、时序特征构建模块和模型预测与训练。在序列嵌入阶段，通过可学习的嵌入层编码项目的语义特征及细粒度时间间隔信息，为后续建模提供基础表示；在时序特征构建中，利用多头注意力权重捕获跨时间步的全局依赖模式，生成具有时序感知的上下文表示。随后，通过堆叠时间卷积神经网络层，采用扩张因果卷积提取局部时序模式，并通过残差连接实现层次化特征聚合；最后基于注意力权重与时序特征的动态融合结果，构建交叉熵损失函数得到下一项目预测概率分布。

3.2 模型目标

设用户集合为 U ，项目集合为 I 。在时间感知的序列推荐场景中，对于每个用户 $u \in U$ ，其行为序列表示为 $S^u = (S_1^u, S_2^u, \dots, S_{|S^u|}^u)$ ，其中 $S_t^u \in I$ 表示用户在时间点 t 交互的项目；对应的时间序列为 $T^u = (T_1^u, T_2^u, \dots, T_{|T^u|}^u)$ ，记录每个行为发生的时间戳。在时间步 t ，序列推荐模型通过分析用户在前 t 个时间点上的交互项目及其相应的时间间隔信息 r_{ij}^u ，预测用户在接下来的时间步中可能的交互项目。因此，模型的输入是 $(S_1^u, S_2^u, \dots, S_{|S^u|-1}^u)$ 以及项目对之间的时间间隔矩阵 $\mathbf{R}^u$ ，其中 $\mathbf{R}^u \in \mathbb{N}^{(|S^u|-1) \times (|S^u|-1)}$ 。模型的预期输出是每个时刻的下一个项目的预测，表现为： $(S_2^u, S_3^u, \dots, S_{|S^u|}^u)$ 。

3.3 模型结构

本章所提出的模型框架如图 3-1 所示。序列嵌入模块利用可学习的位置编码和时间间隔矩阵，联合建模项目的语义特征与细粒度时序上下文信息。为捕捉序列元素的全局交互依赖，引入时间间隔感知自注意力机制，并结合扩张因果卷积神经网络提取局部时序模式。在此基础上，模型进一步融合注意力权重与时序特征，并采用负采样策略构建交叉熵损失函数，以预测下一交互项目的概率分布。

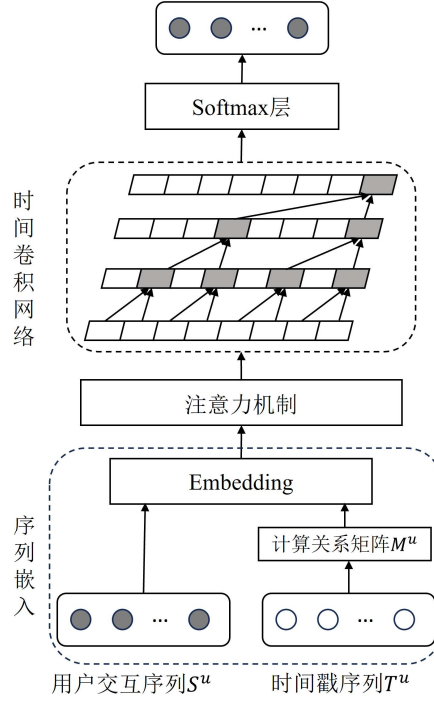


图 3-1 TCSRec 模型的总体框架

Fig.3-1 The overall architecture of the TCSRec model

3.3.1 序列嵌入模块

由于不同用户的交互序列长度不同，为了方便批处理和训练稳定，所有用户的序列都被裁剪或补齐为长度 n ，即如果交互序列长度超过 n ，则保留最近的 n 个交互；如果交互序列长度不足，则在序列左侧补充占位项目，以保持序列对齐。将序列 $(S_1^u, S_2^u, \dots, S_{|S^u|-1}^u)$ 转换为一个固定长度的序列 $s = (s_1, s_2, \dots, s_n)$ ，其中 n 表示设定的最大长度。同样，将时间序列 $T^u = (T_1^u, T_2^u, \dots, T_{|T^u|-1}^u)$ 转换成一个固定长度的时间序列 $t = (t_1, t_2, \dots, t_n)$ 。若时间序列长度大于 n ，则仅保留前 n 个时间戳。否则，用第一个时间戳 $T_{s_1}^u$ 进行填充（前提是 s_1 不是填充项目）。在交互序列中，时间间隔被看作是项目之间的关系。由于不同用户的交互频率存在差异，一些用户可能更频繁地进行交互，而另一些则较少。因此，本研究的重点是分析每个用户在其交互序列中的相对时间间隔。为了解决用户交互频率差异的问题，本文通过使用每个用户序列中最短的非零时间间隔来对所有时间间隔进行归一化处理，从而生成个性化的时间间隔表示，帮助更准确地建模用户的行为。具体来说，给定用户

u 的固定长度时间序列 $t = (t_1, t_2, \dots, t_n)$ ，项目 i 和项目 j 之间的时间间隔为 $|t_i - t_j|$ ，用户 u 的时间间隔集合为 $\mathbf{R}^u$ 。不同用户的活跃度差异极大：高活跃用户一天可能交互几十次，而低活跃用户可能几天才访问一次。因此，单纯用绝对时间差描述用户行为节奏是不合理的。本文使用用户级时间归一化策略：计算用户 u 的非零最小时间间隔为 $r_{\min}^u = \min(\mathbf{R}^u)$ ，然后对该用户行为序列中所有项目之间的时间间隔进行归一化处理，即将每个时间间隔除以该最小时间间隔，即 $r_{ij}^u = \left\lfloor \frac{|t_i - t_j|}{r_{\min}^u} \right\rfloor$ ，其中 r_{ij}^u 是一个自然数。最后构成用户 u 的时间间隔矩阵为：

$$\mathbf{M}^u = \begin{bmatrix} r_{11}^u & r_{12}^u & \dots & r_{1n}^u \\ r_{21}^u & r_{22}^u & \dots & r_{2n}^u \\ \dots & \dots & \dots & \dots \\ r_{n1}^u & r_{n2}^u & \dots & r_{nn}^u \end{bmatrix} \quad (3-1)$$

设定两个项目之间的最大相对时间间隔截断阈值为 k （大于 k 的均设置为 k ），采用截断处理既能避免稀疏关系编码的问题，又能提升模型对训练数据中未出现过的时间间隔的泛化能力。具体实现通过截断关系矩阵 $\mathbf{M}_{clipped}^u = clip(\mathbf{M}^u)$ 完成，其中对矩阵中每个元素执行截断操作 $r_{ij}^u = \min(k, r_{ij}^u)$ 。接着，构建一个项目嵌入矩阵 $\mathbf{M}^I \in \mathbb{R}^{I \times d}$ ，其中 d 是潜在维度。对于填充项，使用一个恒定的零向量 $\mathbf{0}$ 作为其嵌入。通过嵌入查找操作，检索前 n 项的嵌入，并将它们拼接在一起，得到一个矩阵 $\mathbf{E}^I \in \mathbb{R}^{n \times d}$ ：

$$\mathbf{E}^I = \begin{bmatrix} m_{s_1} \\ m_{s_2} \\ \dots \\ m_{s_n} \end{bmatrix} \quad (3-2)$$

同时构建两个不同的可学习的位置嵌入矩阵 $\mathbf{P}^K \in \mathbb{R}^{k \times d}$ 和 $\mathbf{P}^V \in \mathbb{R}^{k \times d}$ ，其中 k 表示最大时间间隔。通过使用 $\mathbf{M}^u$ 中的元素作为索引，分别从 $\mathbf{P}^K$ 和 $\mathbf{P}^V$ 中获取对应的时间间隔编码向量。可以将 $\mathbf{P}^K$ 和 $\mathbf{P}^V$ 中的每一行向量看做是自注意力机制中

的 key 和 value 的偏置向量。可获得其嵌入为 $\mathbf{E}_K^P \in \mathbb{R}^{n \times d}$ 和 $\mathbf{E}_V^P \in \mathbb{R}^{n \times d}$ 。与位置嵌入相似，对于自注意力机制中的 key 和 value，相对时间间隔嵌入矩阵为 $\mathbf{M}_K^R \in \mathbb{R}^{k \times d}, \mathbf{M}_V^R \in \mathbb{R}^{k \times d}$ ，获得截断关系矩阵 $\mathbf{M}_{clipped}^u$ 后，得到嵌入矩阵为 $\mathbf{E}_K^R \in \mathbb{R}^{n \times n \times d}$ 和 $\mathbf{E}_V^R \in \mathbb{R}^{n \times n \times d}$ ：

$$\mathbf{E}_K^R = \begin{bmatrix} r_{11}^k & r_{12}^k & \cdots & r_{1n}^k \\ r_{21}^k & r_{22}^k & \cdots & r_{2n}^k \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1}^k & r_{n2}^k & \cdots & r_{nn}^k \end{bmatrix} \mathbf{E}_V^R = \begin{bmatrix} r_{11}^v & r_{12}^v & \cdots & r_{1n}^v \\ r_{21}^v & r_{22}^v & \cdots & r_{2n}^v \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1}^v & r_{n2}^v & \cdots & r_{nn}^v \end{bmatrix} \quad (3-3)$$

3.3.2 时序特征构建模块

TCSRec 利用注意力机制来衡量用户行为序列中每个元素与用户信息之间的相关性，从而提升关键行为的权重。此外，注意力机制的引入有助于时间卷积网络识别并建模用户长期行为中的周期性模式。对于每个包含 n 个项目的输入序列 $\mathbf{E}^I = (m_{s_1}, m_{s_2}, \dots, m_{s_n})$ ，其中 $m_{s_i} \in \mathbb{R}^d$ ，计算出一个新的序列 $\mathbf{Z} = (z_1, z_2, \dots, z_n)$ ，其中 $z_i \in \mathbb{R}^d$ 。每个输出元素 z_i 是通过对输入元素进行线性变换后，结合关系或位置嵌入的加权求和计算得到的。计算方式如下：

$$z_i = \sum_{j=1}^n \alpha_{ij} (m_{s_i} W^V + r_{ij}^v + p_j^v) \quad (3-4)$$

其中 $W^V \in \mathbb{R}^{d \times d}$ 是 value 的输入投影。注意力权重 α_{ij} 通过 softmax 函数计算：

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^n \exp e_{ik}} \quad (3-5)$$

公式中相关性分数 e_{ij} 使用项目输入、时间间隔嵌入和位置嵌入的来计算：

$$e_{ij} = \frac{m_{s_i} W^Q (m_{s_j} W^K + r_{ij}^k + p_j^k)^T}{\sqrt{d}} \quad (3-6)$$

其中 $W^Q \in \mathbb{R}^{d \times d}$ ， $W^K \in \mathbb{R}^{d \times d}$ 为 Q、K 的线性变换矩阵。在获得注意力机制的输出后，采用时间卷积神经网络对序列的时间依赖关系进行建模，时间卷积网络通过

使用扩张因果卷积来处理用户的行为序列。在该网络中，扩张因果卷积能够通过增大扩张因子来扩大卷积核的感受野，从而捕捉更长时间跨度的行为特征。这种方式不仅提升了网络在长期依赖建模中的能力，而且通过减少参数量，提高了模型的效率。通过这种设计，时间卷积网络能够有效学习到用户行为中的长期模式。时间卷积网络中的卷积核定义如下：

$$f = [w_0, \dots, w_k, \dots, w_{K-1}] \in \mathbf{R}^{d \times K} \quad (3-7)$$

其中 K 是卷积核的大小， w_k 是卷积核中的权重向量。用户在第 t 个时间步的卷积输出 $F(z_t)$ 计算为：

$$F(z_t) = \text{ReLU} \left(\sum_{k=0}^{K-1} w_k \odot z_{t-k-d} \right) \quad (3-8)$$

其中， ReLU 为激活函数， $d \in [1, 2, 4, \dots, 2^n]$ 为时间卷积网络的扩张系数。符号 $\odot$ 表示逐元素相乘。为了增强模型的表达能力，并缓解多层卷积层叠加导致的梯度消失问题，在卷积层使用残差结构。在第 m 层时间卷积网络应用残差模块后，得到的输出可以表示为：

$$H_m(z_t) = F_m(z_t) + z_t \quad (3-9)$$

其中 $F_m(z_t)$ 是第 m 层的卷积输出， z_t 为卷积层的输入。

3.3.3 模型预测与训练

在时间卷积网络处理完注意力机制的输出后，得到一个新的特征序列 $H_M(Z) = [h_1, h_2, \dots, h_T]$ ，随后使用全连接层将这些特征映射到项目空间，并计算用户对每个项目的预测得分 $y_{u,i}$ ：

$$R_{u,i} = h_T \cdot \mathbf{M}_i^T \quad (3-10)$$

其中 $\mathbf{M}_i^T \in \mathbb{R}^d$ 是项目 i 的嵌入表示， h_T 是时间卷积网络的最后一层输出。

在序列推荐任务中，模型的主要目标是根据用户的历史行为序列，预测用户

下一次可能的交互项目。为了实现这一目标，模型需要计算用户对每个候选项目的偏好得分，并通过优化损失函数来调整模型参数，使得预测结果尽可能接近真实数据分布。使用 Adam 优化模型参数，模型训练损失函数采用交叉熵损失函数：

$$L = \frac{1}{N} \sum_{\substack{u \in K \\ i \in R}} [y_{u,i} \log_2 \hat{y}_{u,i} + (1 - y_{u,i}) \log_2 (1 - \hat{y}_{u,i})] \quad (3-11)$$

其中， $y_{u,i} \in \{0,1\}$ 表示用户 u 对项目 i 的真实交互标签（1 表示交互过，0 表示未交互）； $\hat{y}_{u,i}$ 是模型输出的概率值，表示用户 u 对项目 i 的预测概率，通过 Sigmoid 函数归一化到 $[0, 1]$ 范围内的概率值， $\hat{y}_{u,i} = \sigma(R_{u,i})$ 。

3.4 实验与分析

为了评价本章提出的模型 TCSRec 的性能，文中介绍了实验设置，在两个真实数据集上分别开展多组实验，并呈现实验结果。实验内容包括对所提出方法与现有先进基线方法在序列推荐任务中性能对比实验，以验证其有效性，同时通过超参数实验探讨关键参数对模型性能的具体影响。

3.4.1 实验设置

(1) **数据集**。实验使用两个真实世界的数据集。MovieLens 数据集是一个广泛用于评估协同过滤算法的基准数据集。本文使用包含 100 万用户评分的版本 (MovieLens-1M)。Amazon 数据集是大规模商品评论数据集，实验选取其中的 Movies&TV 数据集，这个数据集是高度稀疏的。

(2) **评价指标**。本文采用两种常见的 Top-N 指标，即 Hit Rate@K 和 NDCG@K 来评估推荐性能，实验中 K 取值为 10。对于每个用户 u ，随机抽取 100 个负样本项目，并将这些项目与真实项目一起进行排序。Hit@10 计算前 10 个项目中真实项目的比例，计算公式如下：

$$HR@K = \frac{1}{N} \sum_{i=1}^N hits(i)@K \quad (3-12)$$

其中， N 代表所有用户的数量， $hits(i)@K$ 是一个二元指示函数，表示对于第 i 个

用户，如果其推荐列表中包含了至少一个真实交互项，则取值为 1，否则为 0。

NDCG@10 考虑了位置因素，并对较高位置的项目赋予更高的权重，计算公式如下：

$$NDCG@K = \frac{NCG@K}{IDCG@K} \quad (3-13)$$

$$DCG = \sum_{i=1}^n \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (3-14)$$

$$IDCG = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)} \quad (3-15)$$

其中， DCG 是折损累积增益（Discounted Cumulative Gain）， rel_i 是第 i 个结果的相关性得分， n 是结果的总数。 $IDCG$ 是理想折损累积增益（Ideal Discounted Cumulative Gain），表示在理想排序下的折损累积增益，其中 $|REL|$ 是相关结果的总数。 $NDCG$ 通过将实际的 DCG 与理想的 $IDCG$ 进行比较，得到一个归一化的值，范围在 0 到 1 之间，值越高表示排序质量越好。

(3) 对比方法。本文将 TCSRec 与以下方法进行比较。这些方法包括经典的通用推荐方法（例如 POP，BPR），这些方法不考虑顺序模式；基于一阶马尔可夫链的推荐方法（例如 FPMC，TransRec）；以及基于神经网络的推荐方法（例如 GRU4Rec+，Caser，MARank，TiSASRec）。

- POP^[22]：所有项目根据其在所有用户训练集中的受欢迎程度进行排名，并通过计数操作次数来计算受欢迎程度。

- BPR^[12]：一个贝叶斯个性化排序模型，BPR 通过最大化后验概率对项目进行排序，从而生成个性化推荐。该模型仅依赖于用户与项目之间的隐性反馈信息进行学习。

- FPMC^[16]：该方法结合了矩阵分解和一阶马尔可夫链，分别捕获了长期偏好和动态转移。

- TransRec^[15]：此方法将每个用户建模为从一项到另一项的翻译向量。这是捕获转换的一阶方法。

- GRU4Rec+^[22]：是一种针对用户操作序列进行建模的会话推荐算法，与 GRU4Rec 相比，它采用了不同的损失函数和采样策略，从而在 GRU4Rec 的基础上实现了显著的改进。

- Caser^[26]: 将一系列最近的项目嵌入到时间和潜在空间的“图像”中。该方法可以捕获考虑 L 个最近项的高阶马尔可夫链。

- MARank^[83]: 该方法考虑最近的项目，并应用多序列注意来捕获单个和联合级别的项目依赖关系。

- TiSASRec^[37]: 将时间间隔信息作为重要特征，通过自注意力机制学习不同项目、绝对位置及时间间隔的权重，并进行推荐。

(4) **参数设置**。为进行公平的比较，对于所有模型的设置进行了统一，设置潜在向量维度为 50，正则化系数为 0.05，批处理大小为 128，dropout 设置为 0.2，采用 Adam 梯度下降法进行损失函数优化。最大时间间隔范围 {1, 16, 64, 256, 512, 1024, 2048} 分别进行实验。

3.4.2 性能比较

本章提出的 TCSRec 和相关对比方法在 MovieLens-1M 和 Amazon Movies&TV 数据集上的性能如表 3-1 所示，关键分析如下：

(1) 在基线方法中，MARank 凭借其捕捉个体与联合层面用户交互的出色能力，以及对稀疏数据集的良好处理，展现出了与其他基线相比的卓越性能。然而，当面对密集数据集时，基于神经网络的方法（例如 GRU4Rec+、Caser）则展现出了更为明显的优势，这得益于它们捕捉长期序列模式的能力，这对于密集数据集尤为重要。

(2) 尽管基于马尔可夫链的方法（如 FPMC、TransRec）在动态项目转换方面表现出色，但在处理密集数据集时，其性能往往不如基于神经网络的方法。这可能是因为马尔可夫链方法更多地关注于项目间的短期转移，而忽视了长期用户偏好和复杂动态模式的捕捉。

(3) TCSRec 模型在 NDCG@10 和 HR@10 两个评估指标上均超越了其他所有对比方法。与 TiSASRec 相比，本文提出的方法通过引入时间间隔感知自注意力机制与时间卷积神经网络，构建了一个多层次时序建模框架，从而更有效地捕捉和利用时间上下文信息。

表 3-1 性能比较

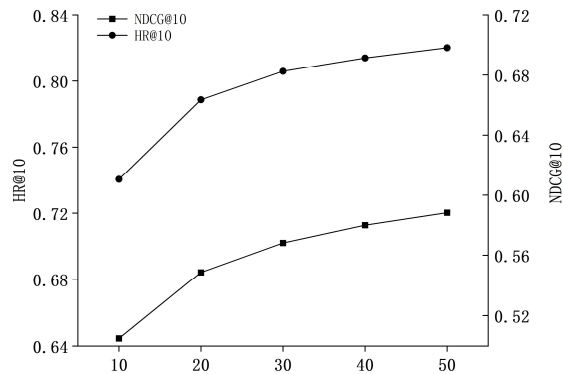
Table 3-1 Performance comparison

模型	MovieLens-1M		Amazon Movies&TV	
	NDCG@10	HR@10	NDCG@10	HR@10
POP	0.2389	0.4386	0.2723	0.4576
BPR	0.3421	0.5952	0.3482	0.5537
FPMC	0.3917	0.6182	0.3376	0.5104
TransRec	0.2488	0.4478	0.2891	0.4793
GRU4Rec+	0.4334	0.6522	0.2987	0.4848
Caser	0.5011	0.7517	0.3212	0.5015
MARank	0.5199	0.7652	0.4568	0.6584
TiSASRec	0.5706	0.8038	0.4974	0.7125
TCSRec	0.5883	0.8202	0.5316	0.7488

3.4.3 超参实验分析

本小节重点探讨模型关键超参数对推荐性能的影响，主要包括：最大序列长度 l 的影响和最大时间间隔 k 的影响。实验过程中，除 l 和 k 之外，其余超参数均保持不变，以确保结果的可对比性。

(1) 最大序列长度 l 的影响：为了探讨最大序列长度 l 对推荐性能的影响，如图 3-2 展示了在 MovieLens 数据集上最大长度 l 从 10 到 50 的 NDCG@10 和 HR@10。实验结果表明，随着 l 的增加，模型的推荐性能逐渐提升。这表明，考虑更长的用户行为序列能够帮助模型更好地理解用户的兴趣和偏好，从而提升推荐效果。然而， l 的增加也会带来模型复杂度的提升，导致训练时间和计算资源消耗的增加。

图 3-2 最大长度 l 对 TCSRec 的影响Fig.3-2 The impact of maximum length l on TCSRec

(2) 最大时间间隔 k 的影响：为了进一步分析最大时间间隔 k 对推荐性能的影响，如图 3-3 展示了在 MovieLens 数据集上，不同最大时间间隔 k 下的 NDCG@10 和 HR@10。实验结果表明，TCSRec 在 $k=256$ 时达到最佳性能，而过大或过小的 k 都会降低推荐效果。 k 值的大小反映了模型对用户行为短期记忆和长期趋势的重视程度。 k 值过大可能导致模型忽略短期用户偏好，降低推荐的时效性，并增加训练成本；而 k 值过小可能削弱对长期行为模式的建模能力。

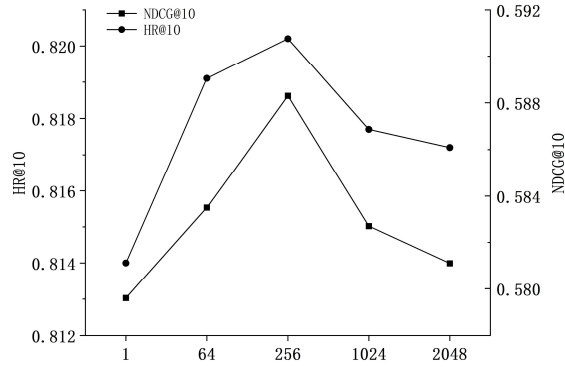


图 3-3 最大时间间隔 k 对 TCSRec 的影响

Fig.3-3 The impact of maximum time interval k on TCSRec

3.5 本章小结

本章围绕时间上下文增强的单域序列推荐算法展开研究，并提出了一种融合时间上下文信息的单域序列推荐方法 TCSRec。本章首先介绍了模型的研究背景，并给出了序列推荐任务的形式化定义。随后，对 TCSRec 的整体框架进行了概述，该模型主要包括序列嵌入模块、时序特征构建模块以及模型的训练与预测。其中，时序特征构建模块作为核心部分，充分考虑了交互事件之间的时间间隔信息，对用户的跨时间步交互依赖关系进行了细粒度捕捉与建模。值得一提的是，TCSRec 结合多头自注意力和多层扩张因果卷积网络，分别捕捉序列中的全局交互特征和局部时间依赖模式，实现了对用户动态兴趣的多层次特征表达。在 MovieLens-1M 和 Amazon Movies&TV 数据集上的对比实验结果表明，所提出的 TCSRec 在推荐准确率和排序效果等指标上均取得了显著优势，验证了所提方法的有效性。

第4章 社交增强时间感知的跨域序列推荐

虽然本文第3章提出的单域序列推荐方法 TCSRec 在结合时间信息以建模用户行为方面取得了较好的效果,但在实际应用中,用户交互数据稀疏且用户偏好易受社交关系影响,限制了推荐性能。为此,本章进一步提出了社交增强时间感知的跨域序列推荐方法来提高推荐模型的预测性能。4.1节分析了现有序列推荐方法存在的问题,并提出对应的解决方案。4.2节对模型基础知识进行介绍,并给出跨域序列推荐问题的形式化定义。4.3节首先介绍了 STRec 的整体框架,即由初始嵌入模块、图神经网络模块和序列注意力编码器模块和社交感知注意力网络模块这三大模块组成,进一步详细介绍这些模块的功能和工作原理。4.4节将本章提出的模型与现有对比方法进行比较。

4.1 研究背景

序列推荐模型以其动态的建模能力和对序列模式的敏锐捕捉,使得推荐结果更加符合用户的实际需求和兴趣变化。因此,近年来序列推荐模型受到了广泛而深入的研究,并在电影、音乐推荐、电子商务等多个领域得到广泛应用^[84,85]。近年来,研究人员不断提升序列推荐系统的准确性。他们采用了如循环神经网络^[86]和卷积神经网络^[20]等先进的神经网络架构,以便更好地处理序列数据。这些架构能够理解用户行为的历史序列,从而提高推荐的准确性和个性化水平。随着自注意力机制的发展,尤其是在 Transformer 模型中的应用,使得 Transformer 的注意力机制序列模型的设计^[87]增加。自注意力机制通过捕捉用户行为之间的长期依赖性,为用户行为的序列动态提供深入和全面的理解。研究人员通过不断优化这些模型,旨在为用户提供更加相关和及时的内容推荐,增强用户体验。尽管有了这些进步,挑战仍然存在,尤其是传统的序列推荐主要关注单个序列中项目的转换模式,缺乏足够的表达能力来对短期用户兴趣进行建模,这对于缺乏历史数据的冷启动用户来说尤为明显。

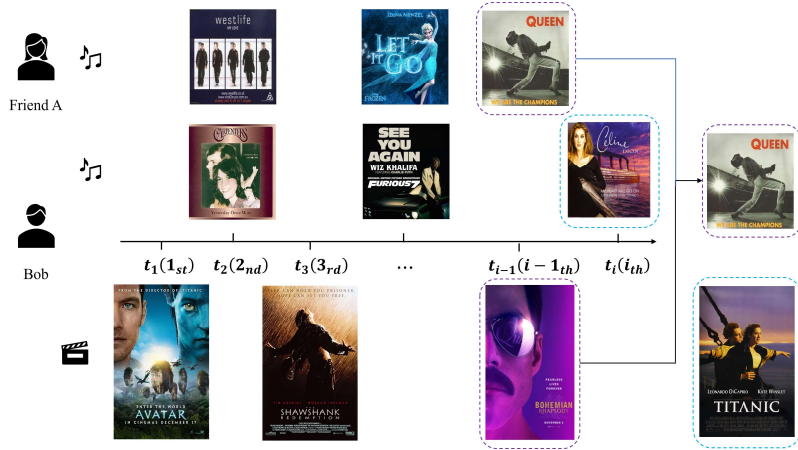


图 4-1 电影与音乐领域用户行为关联示例

Fig.4-1 An example of user behavior association between movie and music domains

跨域序列推荐被研究者提出解决上述问题,跨域序列推荐基于这样一个假设:一个领域中的用户行为序列可能会影响他们在另一个领域的潜在兴趣和偏好。例如,如图 4-1 所示,用户在观看了电影《波希米亚狂想曲》后,可能倾向于在音乐平台上听与电影相关的 Queen 乐队的音乐,如同名的《波希米亚狂想曲》。这个例子说明了 CDSR 的概念,即电影行业的用户观看行为激发了他们在音乐领域的兴趣,从而反映了不同领域中用户偏好的相互关联性。

为解决上述问题,本文提出了一个用于跨域序列推荐的时间和社交增强模型 (STRec),此模型能同时考虑序列内和序列间的项目关系,从而能够学习用户在单域和跨域的项目偏好,并在建模过程中同时考虑社交结构和个人偏好的影响,从而提高推荐性能。具体而言,该模型由三个部分组成:首先,通过时间戳嵌入模块和项目嵌入模块,将时间戳信息和项目信息转换为嵌入向量表示,在这个过程中能够建模用户的长期偏好。其次,基于上一步的嵌入信息,模型使用图神经网络模块挖掘序列间项目的协同特征。随后通过一个序列注意力编码器捕获交互序列内的项目顺序特征,同时从不同上下文中的时间戳捕获潜在的依赖关系,从而更准确地获得用户的短期兴趣。最后,基于用户的社交网络信息,利用一个社交关系注意力网络,计算用户之间的社交影响力,通过将用户的社交邻居信息考虑进来,最后得到用户可能交互的下一项项目的概率。该模型通过结合时间感知和用户社交关系,显著提高了推荐的准确性和个性化程度,同时缓解了数据稀疏问题。

4.2 模型目标

给定目标用户 $u \in U$ ，考虑一个一般的跨域序列推荐场景，其中每个用户的交互序列涉及两个领域，即领域 A 和领域 B。 $\mathcal{S}_u$ 表示用户 u 的交互序列集， $(S^A, S^B, S) \in \mathcal{S}_u$ ，其中用户 u 在领域 A 的交互历史表示为 $S^A = (A_1, A_2, \dots, A_{|S^A|})$ 、在领域 B 的交互历史表示为 $S^B = (B_1, B_2, \dots, B_{|S^B|})$ 。其中每个序列详细记录了交互的时间顺序。跨领域交互序列 $S = (A_1, B_1, B_2, A_2, \dots, A_{|S^A|}, \dots, B_{|S^B|})$ 是通过按时间先后顺序合并用户 u 在领域 A 和 B 的交互序列 S^A 和 S^B 得到的。领域 A 和 B 中所有交互过的项目集合分别用符号 $\mathbb{A}$ 和 $\mathbb{B}$ 表示。当用户 u 与这些领域中的某个项目交互时，则记为 $A \in \mathbb{A}$ 或者 $B \in \mathbb{B}$ 。其中 $|S^A|$ 、 $|S^B|$ 分别是用户在两个不同领域中交互的项目总数。 T^A 、 T^B 是用户在 A 和 B 两个不同领域内与项目交互的时间戳序列。 T 表示完整的时间序列。除了用户的交互序列 $\mathcal{S}_u$ ，还存在用户 u 的社会关系 $F_u = \{(u, v) | u, v \in U\}$ ，定义为一组用户对 (u, v) ，其中 u 和 v 均属于用户集合 U ，用户对 (u, v) 表示用户 u 与用户 v 之间存在社交连接。跨领域序列推荐的主要目标是通过评估未来交互的可能性，预测用户的行为。通过结合用户的历史交互序列 $(S^A, S^B, S)_u$ 和社会关系 $F_u \in U$ ，CDSR 的目标是预测用户在两个领域中与潜在项目交互的概率。该方法着重于确定用户最有可能选择进行下一次交互的项目。预测公式如下：

$$P = \max(P^A, P^B) \quad (4-1)$$

其中，领域 A 和 B 中候选项目的概率分别表示为 $P^A \in \mathbb{R}^A, P^B \in \mathbb{R}^B$ ，为了生成下一次推荐，概率最高的候选项目被选择作为下一个推荐项目。

4.3 模型结构

本章所提出的模型的整体框架如图 4-2 所示，主要包括以下四个部分。嵌入初始化组件负责初始化项目嵌入和时间戳嵌入；随后通过应用图神经网络，捕获每个序列中项目之间的高阶依赖关系，从而更好地表示项目交互中产生的协同信号；时间社交感知注意力模块根据每个项目的上下文相关性确定其重要性，并赋予不同的权重；最后，通过联合训练策略，优化跨领域的交叉熵目标函数。

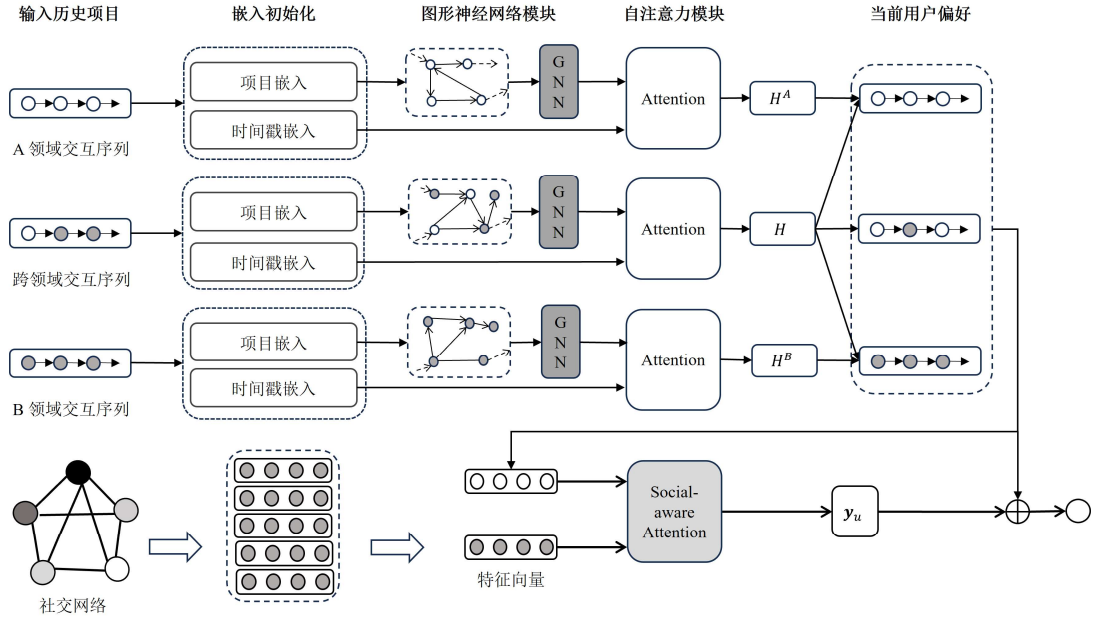


图 4-2 STRec 模型的整体框架

Fig.4-2 The overall architecture of the STRec model

4.3.1 嵌入初始化

项目嵌入框架包含三个参数矩阵。针对特定领域的表示，本文分别使用 $\mathbf{E}^A \in \mathbb{R}^{|A| \times d}$ 和 $\mathbf{E}^B \in \mathbb{R}^{|B| \times d}$ 来表示领域 A 和领域 B 的项目嵌入。此外，联合矩阵 $\mathbf{E} \in \mathbb{R}^{(|A|+|B|) \times d}$ 同时对两个领域中的实体进行封装。公式中的标量 d 表示嵌入空间内的向量维度。

对于用户嵌入，给定一个用户 u ，模块生成其嵌入 $e_{u_0} \in \mathbb{R}^d$ ，使用向量 $q_u \in \mathbb{R}^d$ 来编码每个用户以捕获社交影响。矩阵 $\mathbf{Q}$ 构建为 $\mathbf{Q} = [q_{u_1}, \dots, q_{u_{|U|}}]$ ，其中 $[\cdot]$ 表示连

接操作。为了整合序列信息，引入了可学习的位置嵌入矩阵 $\mathbf{p} \in \mathbb{R}^{l \times d}$ 。将该嵌入添加到项目嵌入矩阵中，交互序列的最大长度用 l 表示。通过这种方式，将项目的序列合并到模型的表示中。

对于用户的时间戳交互序列，时间戳序列被转换为能够保留连续依赖关系的时间嵌入向量。首先，时间戳序列按照以下过程被转化为相对时间间隔序列：

$$T_r = (0, t_2 - t_1, \dots, t_L - t_1) \quad (4-2)$$

其中， L 表示模型可接受的用户交互序列的最大长度。时间戳序列通过填充第一个时间戳 t_1 来对齐（其中，第一个时间戳对应的项目不是占位符）。在实验中，使用 0 表示占位符。目标是将相对时间间隔转换为嵌入向量，从而封装时间信息。将时间间隔均匀划分为不同区间，并为每个区间分配不同的嵌入向量。其次，采用以窗口函数为中心的方法来确定时间间隔的相关性。最大相对时间间隔用 $r_{\max}$ 表示，所有相对时间间隔被截断并缩放至区间 $[0, k]$ ，得到缩放后的相对时间间隔序列 T_s ，公式如下：

$$T_s = (t_s^1, t_s^2, \dots, t_s^L) = \frac{k \times c(T_r, r_{\max})}{r_{\max}} \quad (4-3)$$

其中 $c(x, y) = \begin{cases} x, & x < y \\ y, & x \geq y \end{cases}$ 。对于时间间隔，创建了一个可训练的嵌入矩阵 $\mathbf{M}_T \in \mathbb{R}^{k \times d}$ ，其中 d 为维度大小。接着，使用窗口函数嵌入方法将相对时间间隔序列 T_s 转换为时间嵌入矩阵 $\mathbf{E}_T$ 。矩阵 $\mathbf{E}_T$ 中的元素 e_{t_i} 的计算公式如下： $e_{t_i} = \sum_{j \in [0, k]} w_{t_i, j} e_j$ 。其中权重 $w_{t_i, j}$ 通过窗口函数获得，且 $w_{t_i, j} = F_{\text{win}}(t_i - j)$ 。

4.3.2 图神经网络模块

本研究采用图神经网络精准捕捉序列数据中项目间的关联关系。为强化协同信号的捕获能力，本文对模型结构进行优化，并引入线性聚合器。具体而言，给定有向矩阵 $\mathbf{M}^A$ 、 $\mathbf{M}^B$ 与联合矩阵 $\mathbf{M}$ ，以及初始化嵌入 $\mathbf{E}^A$ 、 $\mathbf{E}^B$ 和 $\mathbf{E}$ ，卷积输出的数学表达如下： $\mathbf{G}_l^A = \text{norm}(\mathbf{M}^A) \mathbf{G}_{l-1}^A$ ， $\mathbf{G}_l^B = \text{norm}(\mathbf{M}^B) \mathbf{G}_{l-1}^B$ ， $\mathbf{G}_l = \text{norm}(\mathbf{M}) \mathbf{G}_{l-1}$ 。其中下标 l 从 0 开始，初始卷积输出为嵌入初始化值 $\mathbf{G}_0^A = \mathbf{E}^A$ ， $\mathbf{G}_0^B = \mathbf{E}^B$ ， $\mathbf{G}_0 = \mathbf{E}$ 。

通过堆叠多层卷积操作，可分别获得不同领域的层级化表征。为进一步挖掘跨层图结构信息，本文对各层输出应用均值函数 $\text{mean}(\cdot)$ 进行信息聚合与平滑，最终得到融合多层特征的项目表征 $\mathbf{G}^A$ 、 $\mathbf{G}^B$ 及联合表征 $\mathbf{G}$ 。

4.3.3 注意力模块

为了精确捕捉序列中项目之间的关系，采用了一个注意力模块来建模交互序列和时间戳。首先，多头自注意力层能够探索交互序列中项目之间的复杂依赖关系，并捕捉在不同上下文中时间戳的潜在依赖关系；其次，逐点前馈层通过非线性变换增强输出序列的表征灵活性。为保持模型简洁性，模型在输入序列中采用项目填充策略，使模型能够适配不同长度的序列，同时提升计算效率。具体流程如下：

$$\begin{aligned}\mathbf{H}^A &= \text{Att}^A(S^A, \mathbf{G}^A, T_S^A) \\ \mathbf{H}^B &= \text{Att}^B(S^B, \mathbf{G}^B, T_S^B) \\ \mathbf{H} &= \text{Att}(S, \mathbf{G}, T_S)\end{aligned}\quad (4-4)$$

其中 $\mathbf{H}^A$ ， $\mathbf{H}^B$ 和 $\mathbf{H}$ 是序列的输出，使用不同参数的多个注意力编码器来分别处理领域特定序列 S^A ， S^B 和跨域序列 S ，以捕捉它们之间的复杂关系。这种方法使得能够针对不同的输入序列进行专门处理，从而更有效地学习序列间的交互信息。通过这种方式，可以更好地适应不同的输入序列，并更精确地捕捉序列之间的关系。

成功捕捉社交影响力是社交序列推荐系统的关键，因为建模用户间动态社交关系能够解释用户为何在特定时间偏好某些项目。然而，现有研究大多未能有效利用社交关系，或以时间不敏感的方式捕获社交影响。为解决这一问题，提出了一种社交感知注意力机制，用于量化在预测用户 u 下一步行为时其每位朋友 v 的重要性。注意力向量的数学表达式如下：

$$\mathbf{H}(u, v, t) = \phi(h_u(t)\mathbf{W}_1 + q_v\mathbf{W}_2 + \mathbf{b}')$$
(4-5)

其中， $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d \times d}$ 与 $\mathbf{b}' \in \mathbb{R}^d$ 为提取有用信息的可学习参数， $h_u(t)$ 表示用户 u 时间步 t 的偏好向量， $\phi(\cdot)$ 采用双曲正切函数实现非线性变换。使用嵌入向量 e_u 作为上下文向量。随后通过 softmax 函数对 $\mathbf{H}(u, v, t)$ 与 e_u 的归一化亲和度计算，得到注意力权重 $\alpha(u, v, t)$ ：

$$\alpha(u, v, t) = \frac{\exp(e_{u*} H^T(u, v, t))}{\sum_{g \in F^u} \exp(e_{u*} H^T(u, g, t))} \quad (4-6)$$

最终，用户 u 在时间 t 的社交影响聚合表征为：

$$y_u(t) = \sum_{v \in F^u} \alpha(u, v, t) \cdot \mathbf{q}_v \quad (4-7)$$

4.3.4 模型预测与训练

基于注意力机制生成的序列输出及用户社交影响力，本研究采用双序列目标优化策略，有效提升模型对单域表示与跨域表示的理解能力。针对单域表示，在处理单域项目预测任务时，采用直接高效的训练方法：构建能够预测用户下一步交互项目的模型。

在领域 A 中，收集用户与项目的交互序列 S^A 。该序列包含用户历史交互的所有项目，并可能添加填充项以维持序列长度一致性。给定 S^A 与目标项目 A_p ，模型需预测用户下一步可能交互的项目 A_{p+1} 。为此，优化编码器的训练策略，其核心职责是从交互序列中提取有效特征，并利用这些特征精准预测下一项目。

$$\mathcal{L}_A = - \sum_{u \in U} \sum_{p=1}^{|A_u|} \log P(A_{p+1} | A_p, \dots, A_1, y_u(t)) \quad (4-8)$$

基于学习到的表征，领域 A 中用户下一步交互项目的概率计算如下：

$$P(A_{p+1} | A_p, \dots, A_1, y_u(t)) = \text{softmax}(\mathbf{H}_u(t) \mathbf{W}_A) \quad (4-9)$$

其中， $\mathbf{H}_u(t)$ 为融合用户自身偏好与社交嵌入 $y_u(t)$ 的综合表征。 $\mathbf{W}_A \in \mathbb{R}^{d \times |A|}$ 为领域 A 的可学习参数矩阵。上述计算过程同等适用于领域 B ，可推导出对应的损失函数 $\mathcal{L}_B$ 。

在处理跨域交互序列时，可以观察到用户行为常表现出复杂多样性。例如，在电影与音乐等不同领域中，用户可能呈现完全不同的偏好模式。此类跨域偏好迁移现象对推荐系统提出了更高要求：传统单域推荐模型难以充分捕捉跨域兴趣转移，亟需新方法应对此类复杂场景。

为构建可同时处理多领域的推荐模型，需重点考虑不同领域间的交互与相互

影响。实际应用中，跨领域交互数据常存在不均衡性：某些领域交互数据丰富，而其他领域则相对稀疏。这种数据失衡可能导致模型训练时过度依赖数据丰富的领域，忽视稀疏领域。为此，需要设计均衡化目标函数，确保模型在不同领域间保持公平性与有效预测能力。跨域损失函数定义为：

$$\mathcal{L}_{cross} = \begin{cases} -\log P_{cross}^A \left(A_{p+1} | [A_1, B_1, \dots, A_p, y_u(t)] \right) \\ -\log P_{cross}^B \left(B_{p+1} | [A_1, B_1, \dots, B_p, y_u(t)] \right) \end{cases} \quad (4-10)$$

在跨域用户交互中，下一步潜在交互项目的概率计算表达式如下：

$$P_{cross}^A = \text{softmax}(\mathbf{H}_p \mathbf{W}_A) \quad (4-11)$$

$$P_{cross}^B = \text{softmax}(\mathbf{H}_p \mathbf{W}_B) \quad (4-12)$$

在上述公式中，将领域 A 和领域 B 的损失函数分解为两个独立且平等的部分进行优化，确保即使另一领域的交互序列包含大量连续项目时，模型仍能保持对另一领域的预测能力。最终优化函数为：

$$L = \lambda(L_A + L_B) + (1 - \lambda)L_{cross} \quad (4-13)$$

其中， λ 为超参数，用于协调单域任务与跨域任务的权重。

4.4 实验与分析

本节详细介绍了几个实验，旨在使用真实世界的数据集评估本章提出模型在预测任务上的性能。实验内容包括性能对比实验，验证所提出模型与现有先进序列推荐方法的性能差异；消融实验，探究 STRec 模型中不同组件对推荐结果的影响；超参数对比实验，分析关键参数对模型性能的影响。

接下来，首先介绍实验设置，再依次对上述研究问题进行回答。

4.4.1 实验设置

(1) **数据集**。实验采用了 Amazon 电子商务数据集，其中包括数百万产品和用户，以及丰富的元数据（如商品类别、评论、描述等）。此外，数据集收集了用户行为数据（浏览、点击、购买等）。在实验设计中，选择了“Food-Kitchen”

两领域中的数据。为了满足跨域序列推荐的特点，首先从数据中提取在两个领域都有过交互的用户，随后剔除了交互次数少于 10 次的用户和商品，以提升模型效果并增强推荐的准确性。用户的最近交互序列被均分至测试集和验证集，未用于验证集和测试集的交互序列构成训练集。

(2) 评价指标。实验采用留一法交叉验证评估推荐系统的性能。为确保评估的公平性，每个验证或测试案例计算 1,000 个得分，其中包括 999 个负样本和 1 个正样本。在评估指标方面，选择了多种度量标准以全面评估推荐系统的有效性，包括：命中率 (HR) @ {1, 5, 10}，归一化折损累积增益 (NDCG) @ {5, 10}，平均倒数排名 (MRR)。

(3) 对比方法。在这项研究中，通过将 STRec 模型与不同类型的先进的基线方法进行对比展示了 STRec 模型的有效性。

- BPRMF^[12]是一种经典的协同过滤算法，在矩阵分解框架下，通过 BPR 损失函数进行优化。

- ItemKNN^[88]分析用户-项目矩阵，计算项目相似度，并利用这些相似度为用户提供高质量的个性化推荐。

- SASRec^[32]使用自注意力机制来建模交互序列。

- TiSASRec^[37]：是一种序列推荐方法，不仅整合了时间间隔信息以提供更准确的预测，还通过自注意力机制有效捕捉用户行为的变化。

- π -Net^[46]这项开创性的研究引入了一种新型的门控递归模块，有效捕捉了每个领域内的序列依赖关系。

- PS-JNet^[47]在 π -Net 的基础上进行扩展，引入了并行跳跃连接方案，实现了不同用户意图的跨领域迁移。

- MIFN^[50]通过构建知识图谱促进跨领域信息的传递，指导不同领域内项目的连接。

(4) 参数设置：实验将小批量大小设置为 256，dropout 率设置为 0.3，嵌入大小为 d 。学习率从 {0.001, 0.0005, 0.0001} 中选择，正则化系数 L_2 从 {0.0001, 0.00005, 0.00001} 中选择。训练总周期数为 100，以确保获得最佳结果。实验选

择 λ 的范围为 0.1 到 0.9，步长为 0.1，图神经网络的深度 L 从 {1, 2, 3, 4} 中选择。所有参数均使用 Adam 优化器进行更新。

4.4.2 性能比较

对比方法和本章提出的 STRec 模型在 Food-Kitchen 数据集上的性能如表 4-1 所示，最佳性能以粗体标出，次优用下划线标出，性能分析如下：

(1) 对于序列推荐基准方法，TiSASRec 和 SASRec 在混合序列推荐数据集上表现良好。同时，这些方法显著优于 BPRMF 和 ItemKNN，表明记录交互序列和时间间隔信息对实现最佳模型性能至关重要。

(2) 在跨域序列推荐基准方法中，表现较好的模型是 PSJNet 和 MIFN，这表明跨领域信息对于提升推荐性能具有积极作用。这也进一步证明，在单领域和跨领域场景下，如何有效建模用户偏好是成功完成 CDSR 任务的关键因素。

(3) 本章提出的 STRec 模型在各种指标上优于所有基准方法，验证了该模型在 CDSR 任务中的有效性。与早期研究结果相比，结合用户社交影响和时间间隔信息对模型性能产生了积极影响。这表明，在单域和跨域场景中，准确模拟用户偏好对于 CDSR 至关重要。

表 4-1 Food-Kitchen 数据集上推荐的性能比较

Table 4-1 performance of strec and baselines on food-kitchen

Method	Food-domain recommendation						Kitchen-domain recommendation					
	MRR	NDCG		HR			MRR	NDCG		HR		
		@5	@10	@1	@5	@10		@5	@10	@1	@5	@10
BPRMF	4.10	3.55	4.03	2.42	4.51	5.95	2.01	1.45	1.85	0.73	2.18	3.43
ItemKNN	3.92	3.51	3.97	2.41	4.59	5.98	1.89	1.28	1.75	0.58	1.99	3.26
GRU4Rec	5.79	5.48	6.13	3.63	7.12	9.11	3.06	2.55	3.10	1.61	3.50	5.22
SASRec	7.30	6.90	7.79	4.73	8.92	11.68	3.79	3.35	3.93	1.92	4.78	6.62
π -Net	7.68	7.32	8.13	5.25	9.25	11.75	3.53	2.98	3.73	1.57	4.34	6.67
PSJ-Net	8.33	8.07	8.77	5.73	10.28	12.45	<u>4.10</u>	<u>3.68</u>	<u>4.32</u>	2.14	<u>5.17</u>	<u>7.15</u>
MIFN	<u>8.55</u>	<u>8.28</u>	<u>9.01</u>	<u>6.02</u>	<u>10.43</u>	<u>12.71</u>	4.09	3.57	4.29	<u>2.21</u>	4.86	7.08
STRec	8.71	8.76	10.04	6.37	11.17	13.94	4.54	3.77	4.74	2.26	5.23	7.63

4.4.3 消融实验分析

为了评估本章提出的 STRec 模型的有效性，本文对 STRec 的不同变体进行

了消融研究，以验证各个模块所带来的优势。本小节设计了 STRec 的两个变体：一是仅使用时间信息的模型（w/o-social），该变体基于时间感知的自注意力跨域序列推荐模型，但未考虑任何社会关系；二是仅使用社交信息的模型（w/o-time），该变体建立在社会感知自注意力跨域序列推荐模型的基础上，但不包含任何时间信息。通过对比完整模型与这两种变体的实验结果，可以深入分析时间信息与社交信息在 STRec 中的具体作用。

实验针对不同变体进行了消融实验，以评估各模块对推荐性能的影响。实验结果如图 4-3 和图 4-4，结果表明将时间间隔信息纳入序列表示可以显著提升性能。这可能是因为时间间隔信息能够有效捕捉用户行为的时间动态变化及偏好的演变。不同的时间间隔可能反映了用户对不同项目的兴趣强度和趋势变化，这有助于模型理解用户行为模式，并做出更准确的推荐。此外，实验还表明，引入社交关系信息能够优化跨域序列表示，进一步提升模型性能，说明社交关系为推荐系统提供了额外的行为线索，有助于增强推荐的准确性。

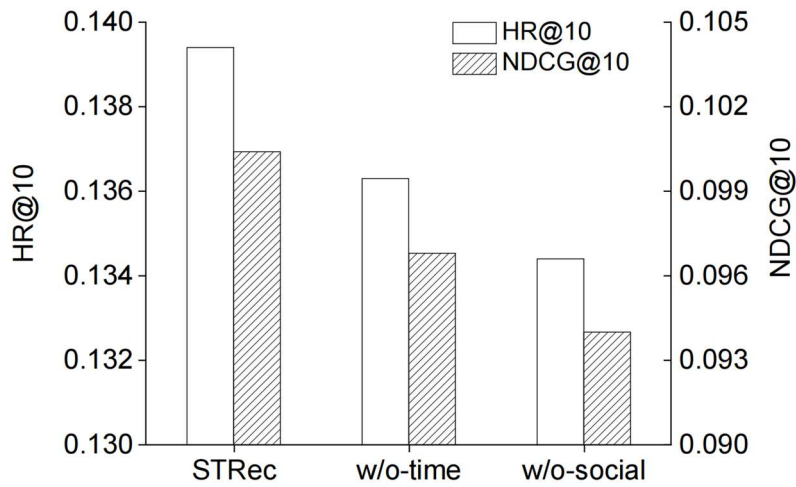


图 4-3 STRec 模型在 Amazon-Food 数据集的消融研究

Fig.4-3 Impact of different components on STRec

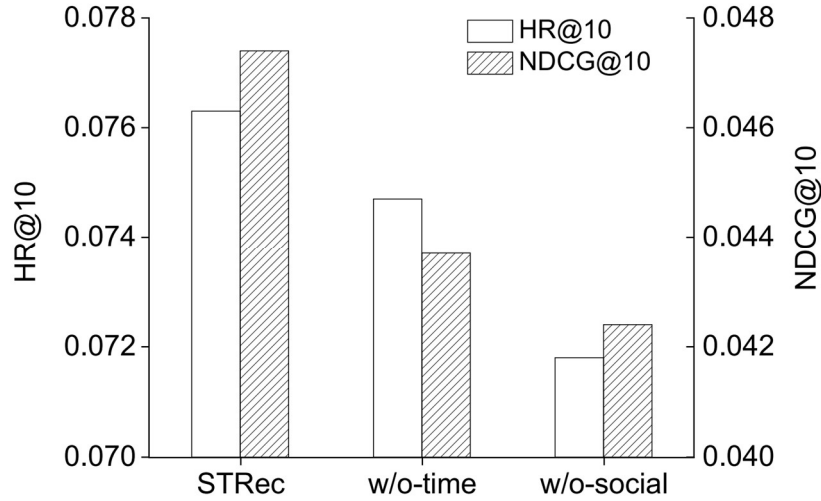


图 4-4 STRec 模型在 Amazon-Kitchen 数据集的消融研究

Fig.4-4 Impact of different components on STRec

4.4.4 超参实验分析

在本小节中，探讨了重要超参数 λ 如何影响模型的性能结果。如图 4-5 和图 4-6 展示了实验结果。实验探索了从 0.1 到 1 的 λ 值，其中 $\lambda=1$ 表示训练仅依赖于推荐任务。实验的结果表明，当 λ 的值在 0.1 到 0.4 之间时，模型的性能较差，而最佳性能出现在 $\lambda=0.7$ 。此外，将 λ 设置为更大的值会导致训练过程中更快的收敛，这表明 $\lambda=0.7$ 在效率和性能之间达到了理想的平衡。

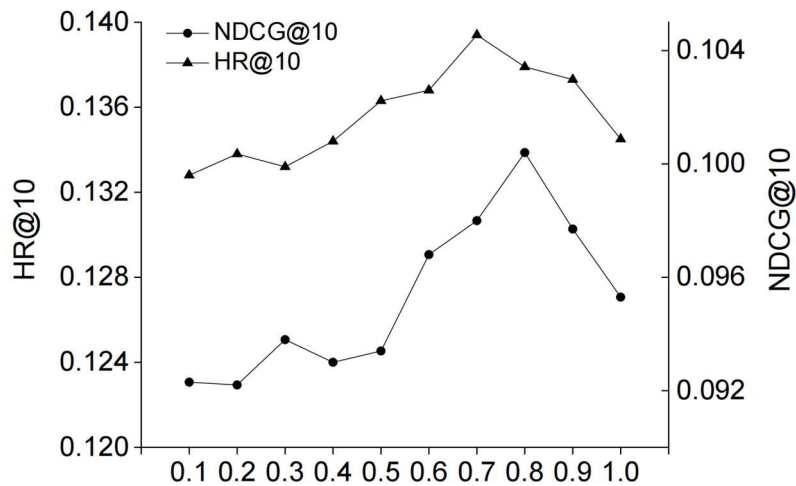
图 4-5 超参数 λ 在 Amazon-Food 数据集中对 STRec 的影响

Fig.4-5 Impact of different components on STRec

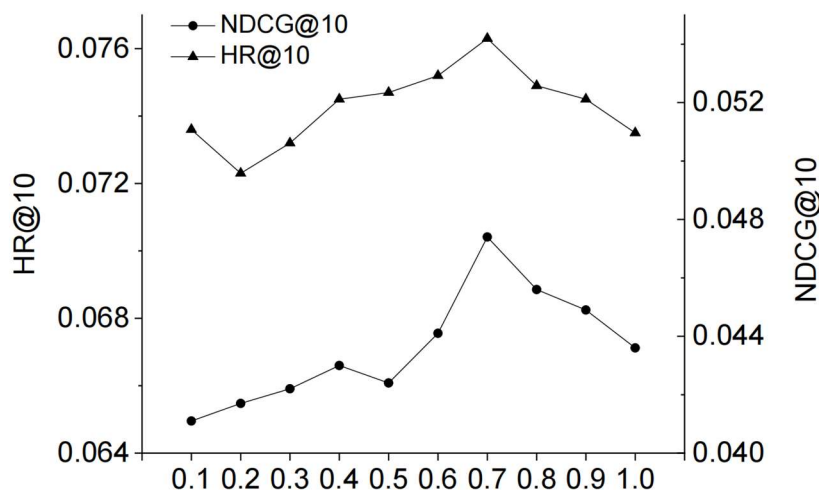
图 4-6 超参数 λ 在 Amazon-Kitchen 数据集中对 STRec 的影响

Fig.4-6 Impact of different components on STRec

4.5 本章小结

本章对社交增强时间感知的跨域序列推荐算法开展研究,并提出一种融合社交关系与时间上下文信息的跨域序列推荐方法 STRec。本章首先介绍了模型的整体框架和关键设计思路,STRec 主要包括嵌入模块、图神经网络模块和注意力模块。其中,图神经网络模块利用图神经网络有效揭示不同领域用户行为序列中项目之间的潜在关联特征;注意力模块充分考虑用户的社交邻居信息以及社交关系的动态演化特征,为跨域用户兴趣建模提供了丰富的上下文信息支撑。在多个跨域推荐场景的真实数据集上进行对比实验,结果表明,所提出的 STRec 能够有效缓解数据稀疏与冷启动问题,并在推荐准确率和多样性等指标上均显著优于现有主流方法,验证了本章所提出方法的有效性与优越性。

第5章 支持隐私保护的端云协同训练框架

在缓解数据稀疏性和提升推荐模型性能方面，第4章提出的社交增强时间感知的跨域序列推荐方法已取得显著效果，但用户数据安全性与隐私保护问题仍然是不可忽视的重要挑战。因此，本章面向推荐系统提出了支持隐私保护的端云协同训练框架来提高用户的数据安全性与隐私保护。5.1节分析了现有隐私保护方法存在的问题，并提出对应的解决方案。5.2节对模型基础知识进行介绍，并给出隐私保护序列推荐问题的形式化定义。5.3节首先介绍了端云协同训练的整体框架，即由云侧模型准备、端侧部署与训练和端云服务这三大模块组成，进一步详细介绍这些模块的功能和工作原理。5.4节将本章提出的模型与现有对比方法进行比较。5.5节展示了端云协同训练框架的原型系统。

5.1 研究背景

传统推荐系统通常依赖集中式数据训练模式，虽然能够有效挖掘用户偏好，但也面临用户隐私泄露和数据安全风险。为此，联邦学习^[89]作为一种新兴的隐私保护技术逐渐受到关注，它允许各设备在不共享原始数据的前提下，基于本地用户交互序列完成模型训练，并通过加密参数传输和全局模型聚合实现联合推荐建模。相比传统数据加密技术，联邦学习在保护数据隐私的同时，有效促进了跨设备、跨平台的数据协同建模，缓解了推荐系统中的数据孤岛问题。然而，现有联邦学习在序列推荐中的应用仍面临诸多挑战：一方面，梯度传输过程中仍存在隐私泄露风险^[90]；另一方面，端侧设备的数据通常存在非独立同分布（Non-Independent and Identically Distributed, Non-IID）和数据量不均衡等问题^[91-92]，导致全局模型训练效果受到显著影响。此外，部分设备因计算资源受限或参与率低，进一步加剧了联邦训练的性能波动，影响推荐效果的稳定性。

为解决上述问题，端云协同训练为隐私保护推荐提供了新的思路。端云协同训练是一种结合端侧设备和云服务器共同计算的分布式训练范式，能够在保证数据隐私安全的前提下，有效利用端侧数据的多样性与实时性，同时弥补云侧计算

资源与用户本地计算能力的差异。现有主流的端云协同训练框架包括 FedML^[70]和 Flower^[93]，它们分别采用 MNN^[94]和 TFlite 作为端侧训练引擎，通过端侧与云侧的协同工作，实现高效的模型训练与用户数据隐私保护。本文基于 FedMNN 框架，针对序列推荐任务构建了一套支持隐私保护的端云协同训练框架，以安卓设备为例，对 FedML+MNN（FedMNN 框架）和 Flower+TFlite（FLTFlite 框架）在端云协同训练功耗进行了对比分析，结果表明：FedMNN 框架在端侧设备上的单轮平均功耗比 FLTFlite 框架降低 32%~51%，具备更优的能效表现。结合 DSSM 和 Deep&Wide 等推荐模型的实验结果，进一步验证了端云协同训练在隐私保护和推荐性能方面的有效性。

5.2 模型目标

随着用户行为数据规模的不断扩大和智能终端设备的广泛普及，用户的交互序列数据呈现出高度分散化和隐私敏感性的特点。这些数据蕴含着用户兴趣动态变化的重要线索，是序列推荐系统构建个性化推荐服务的重要基础。然而，传统序列推荐系统多采用集中式训练模式，将用户数据汇聚至中心服务器进行统一建模，这种模式虽然能够有效捕捉用户偏好特征，但也存在严重的隐私泄露和数据安全隐患。此外，随着数据保护法规的不断完善，用户对个人隐私数据的保护意识日益增强，进一步限制了中心化数据获取的可行性，导致推荐系统面临“数据可用性与隐私保护”的双重困境。为此，联邦学习和端云协同训练逐渐成为隐私保护推荐的重要解决方案。联邦学习通过分布式训练与参数加密聚合，实现用户本地数据不出端、参数加密传输的联合建模，避免了用户行为数据的直接暴露。端云协同训练进一步结合端侧设备的计算能力与云端的全局建模能力，确保模型训练过程既能够兼顾本地个性化特征，又能够融入全局协同信息，从而在保证用户隐私的前提下，构建更精准的序列推荐模型。

基于上述分析，本文定义的隐私保护序列推荐问题如下：在用户交互序列数据分散存储于多个端侧设备的场景下，如何在不直接传输用户行为数据的前提下，基于联邦学习与端云协同训练框架，实现对用户序列数据的动态兴趣建模，并在

全局范围内构建高效、个性化的序列推荐模型，同时兼顾隐私保护与推荐性能的平衡。

5.3 模型结构

本节介绍基于 FedMNN 框架进行端云协同训练的通用流程，并对训练过程涉及到的模块进行详细阐述。此外，以经典推荐算法模型 DSSM 与 Deep&Wide 为例，介绍端云协同场景下，实现不同算法模型的一般流程。基于 FedMNN 的端云协同训练流程分为三个部分：云侧模型准备、端侧部署与训练、端云服务。

5.3.1 云侧模型准备

Tensorflow、Pytorch 和 Caffe 等常见的深度学习框架通常被部署在云侧进行模型训练，不能直接支持端侧设备训练。FedMNN 使用移动神经网络框架 MNN 作为端侧统一训练框架进行端侧训练。从通用性角度出发，为了实现端侧训练，本文将云侧模型转换为端侧可训练的 MNN 模型。模型转换过程如图 5-1 所示，使用开源框架 ONNX 作为中间表示，将不同深度学习框架实现的模型转换为 ONNX 模型，并利用 MNN 提供的模型转换工具将 ONNX 模型转换为 MNN 模型。转换后的 MNN 模型可以直接发送给不同客户端用于端侧训练。

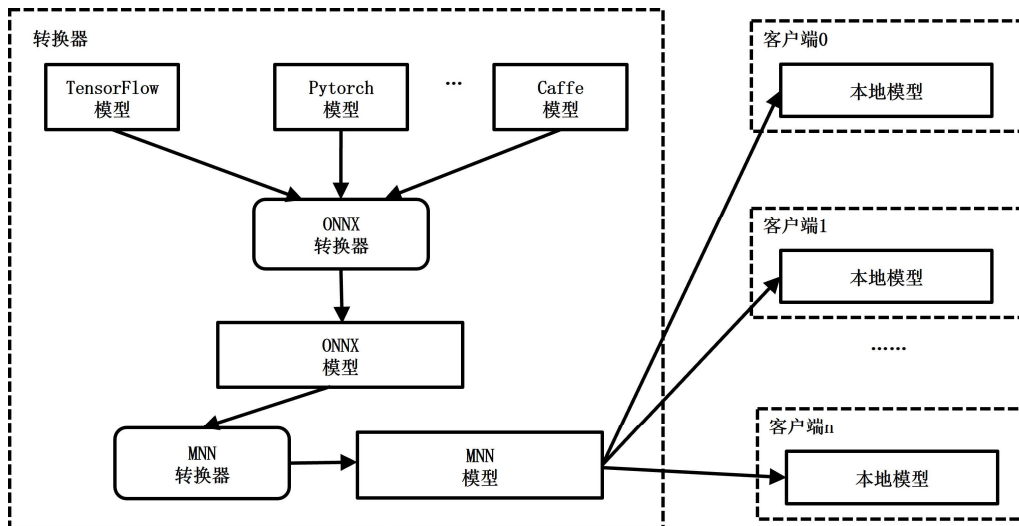


图 5-1 云侧模型转换

Fig.5-1 Model transformation on cloud side

5.3.2 端侧模型部署与训练

端侧模型训练需要完成两个准备阶段，一是模型准备，二是数据准备。在模型准备阶段，任务初始化时，云侧将初始化模型直接下发到端侧。训练过程中，每轮训练结束后，云侧对端侧上传的模型进行聚合，并将聚合模型再次下发给端侧；在数据准备阶段，需要将端侧数据存放到指定路径。此外，基于 MNN 框架进行端侧训练需要针对特定训练数据进行数据集的定义并实现数据读取操作。特此说明，在实验验证阶段，将预处理好的数据上传到手机中，而在实际应用中，用户数据均来自本地，并且在训练过程中无需上传至云侧，在一定程度上保护了用户数据的安全与隐私。

总体来说，端侧训练的流程与云侧训练类似。端侧模型的训练流程如图 5-2 所示。在初始化阶段，端侧获取云侧下发的模型并初始化优化器。然后创建 Executor 对象，给用户配置推理后端、线程数等属性的接口，随后创建 Dataset 以及 DataLoader。在训练阶段，从 DataLoader 中获取训练数据，并执行模型前向操作，然后根据训练任务的不同来计算损失，并执行梯度反向传播，便完成了一个批次的训练。

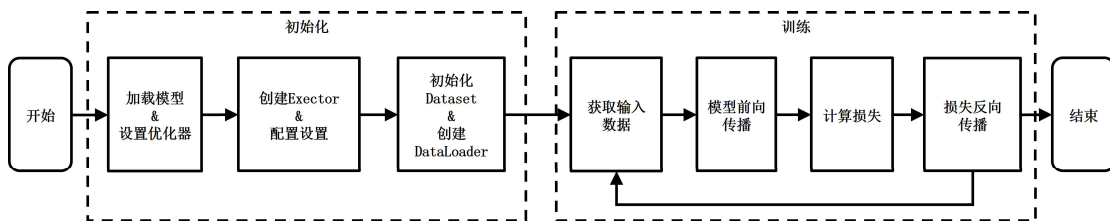


图 5-2 端侧训练

Fig.5-2 Training on end side

5.3.3 端云服务

端云服务是实现端云模型协同优化的重要途径，主要包括三个方面：一是将端侧模型的训练结果反馈给云侧，进行模型的聚合和更新，以提高云侧模型的泛化能力和鲁棒性；二是根据业务需求，将云侧模型按需部署到不同的设备上，实现端云协同服务；三是根据云侧的负载情况对端侧设备进行调度，选择参与训练的设备。

模型聚合与更新。端云协同训练的关键之一在于云侧模型的聚合，将多个端

侧模型的参数或梯度进行加权平均或其他方式的融合，得到一个新的云侧模型。云侧模型聚合服务的整体流程如图 5-3 所示。FedMNN 使用 FedML 作为云侧联邦学习框架，FedML 提供聚合服务，可以选择不同的联邦学习聚合算法，本文选择 FedAvg 算法进行后续实验。其实现原理是将端侧上传的 MNN 模型读取为 Pytorch Tensor 字典进行聚合，并将聚合模型下发至端侧对端侧模型进行更新。

端云通信。FedMNN 完成端云通信依赖 FedML 提供的通信链路，该通信链路基于消息队列遥测传输协议进行消息传输，在数据传输过程中，模型均以 MNN 的形式存储。端云通信链路中的数据以及模型流向如图 5-3 所示。训练初始阶段，由云侧向所有加入训练的端侧设备下发初始化 MNN 模型，端侧设备基于此模型以及本地数据进行训练。训练完成后，所有端侧设备将本地训练的 MNN 模型上传至云侧进行聚合，并等待云侧完成聚合，将聚合后的模型发送给所有端侧设备。

端侧设备调度。训练由云侧发起，并在云侧选择好参与训练的端侧设备。每轮通信中，必须保证通信链路的畅通，当参与训练的任意客户端发生意外断开连接，或者模型上传、下载的过程出现丢失，都将导致本次训练无法继续进行，这也是 FedML 目前的局限性。此外，研究发现目前 FedMNN 中的端侧设备和云服务器还无法根据网络状况、计算能力、数据特征等动态调整通信频率、更新步长、

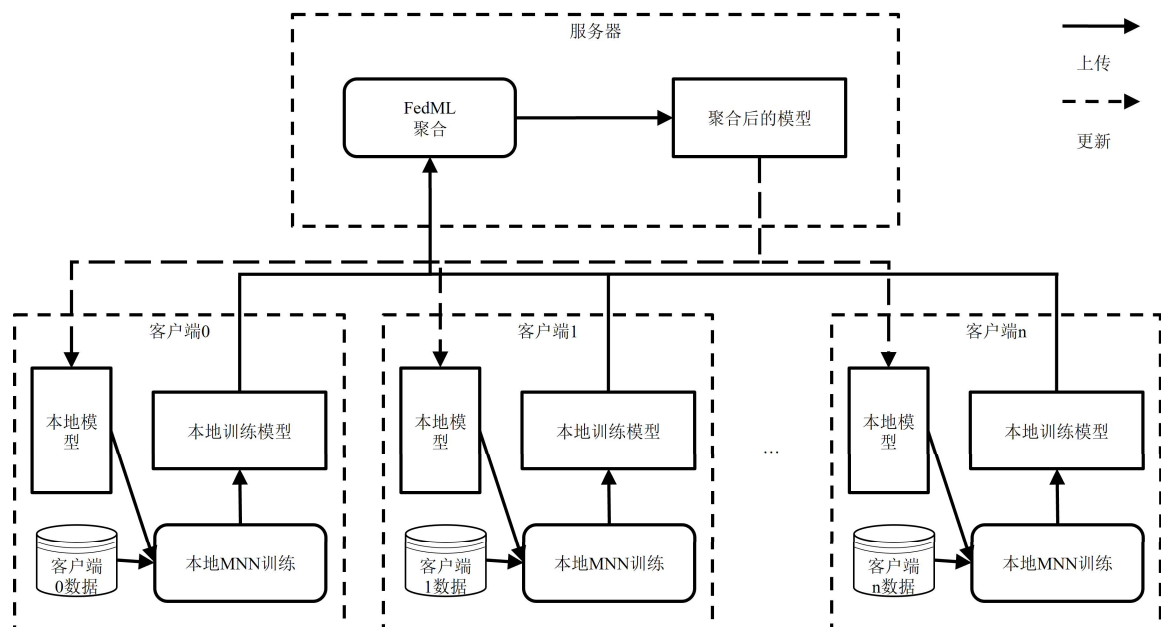


图 5-3 端云模型聚合与更新

Fig.5-3 Cloud-end model aggregation and update

聚合策略等参数，实现自适应的端云协同训练。因此，在适应更大规模的端云协同训练上仍然有很多工作要做。

5.3.4 端云协同训练算法实现

为了说明 FedMNN 在端云协同场景下的应用，本文对 FedMNN 实现端云协同训练的算法进行总结。与传统的 FedAvg 算法相比，在 FedMNN 框架下，需要首先完成模型的转换，将其他模型转换为 MNN 模型进行端云协同训练。此外，FedMNN 在进行端侧设备调度时需要从 MLOps 平台选择参与训练的设备，并将初始化模型发送给所有选中的设备。在训练过程中，需要将端侧上传的模型转换为 Torch tensor 字典进行聚合，再将聚合后的参数存储到 MNN 模型。算法实现的伪代码如算法 1 所示。

算法 1 端云协同训练

输入：初始化全局模型 m_1 ；每个客户端的数据集 $\mathcal{D}_k$ 。

输出：聚合全局模型 m_T 。

- 1: K 个客户端编号为 k ； B 是本地批量大小(batch_size)； E 是客户端训练轮数(epoch)； T 是通信的轮数， η 是学习率。
 - 2: 服务器执行：
 - 3: 将 torch 模型转换为 MNN 格式模型 m_1
 - 4: **For** 每轮 $t=1,2,\dots,T-1$ **do**
 - 5: $S_t \leftarrow$ 选择参与训练的 K 个客户端
 - 6: **For** 每个客户端 $k \in S_t$ **do** // K 个客户端并行处理
 - 7: 发送模型 m_t 到客户端 k
 - 8: $m_{t+1}^k \leftarrow$ 调用客户端更新(k, m_t) // 得到客户端 k 上传的新模型 m_{t+1}^k
 - 9: $w_{t+1}^k \leftarrow m_{t+1}^k$ // 从模型 m_{t+1}^k 中提取模型的参数 w_{t+1}^k
 - 10: $w_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} w_{t+1}^k$ // 模型聚合
 - 11: $m_{t+1} \leftarrow w_{t+1}$ // 将聚合后的模型参数 w_{t+1} 存储到 MNN 模型中，得到 m_{t+1}
 - 12: 客户端更新(k, m_t) // 在客户端 k 上运行
 - 13: 将 $\mathcal{D}_k$ 分成 B 个大小为 B 的批次
-

```

14: 从模型  $m_i$  中提取本地模型的参数  $w_i^k$ 
15:   For 本地第  $i \in (1, E)$  轮 do
16:     For 每个批次  $b \in (1, B)$  do
17:        $w_{i+1}^k \leftarrow w_i^k \leftarrow \eta \nabla l(w_i^k; b)$  //  $\nabla$  为计算梯度,  $l(\cdot)$  为损失函数
18: 保存  $w_{i+1}^k$  为本地 MNN 格式模型  $m_{i+1}^k$ 
19: 返回  $m_{i+1}^k$  至服务器

```

5.4 实验与分析

为了验证 FedMNN 在端云协同场景下端侧设备的功耗以及模型性能, 本文使用安卓手机进行了框架功耗测试。此外, 选取了推荐系统中的经典算法 DSSM 与 Deep&Wide 来验证端云协同训练模型在推荐场景中的性能。

5.4.1 框架功耗测试

为了对比 FedMNN 与 FLTFlite 两种端云协同训练框架在真实端侧设备运行的功耗情况, 本文选择了相同的模型以及联邦学习常用的基准数据集 MNIST 进行了训练功耗测试。MNIST 是一个手写体数字的图片数据集, 该数据集来自美国国家标准与技术研究所。数据集由来自 250 个不同人手写的数字构成, 其中 50% 是高中学生, 50% 来自人口普查局的工作人员。训练集一共包含了 60000 张图像和标签, 而测试集一共包含了 10000 张图像和标签。实验中, Flower 和 FedML 云侧服务均部署在 PC 设备上, 实验过程均使用 FedAvg 算法, 端云通信轮数均设置为 20。使用相同的 Android 手机作为端侧测试设备, 并保证训练时间有间隔, 避免连续实验对手机原始性能产生影响。在训练过程中, 使用 PowerMonitor 功耗仪进行测试, 并使用 PowerTool 工具进行数据分析。

表 5-1 展示了 FLTFlite 与 FedMNN 在端侧设备上的功耗情况。首先, 直接统计训练完成一轮端云通信时的端侧平均功耗, 总体来说, FedMNN 的平均每轮通信功耗比 FLTFlite 低 0~20%。为了更严谨的统计端侧训练过程带来的实际功耗, 本文统计了端云协同训练相对不训练产生的平均功耗增量, 具体操作是统计使用 APP 在不进行训练时的平均功耗, 并用每轮通信的实际功耗减去平均功耗

乘平均每轮训练时间。通过表 5-1 可以发现，使用 FedMNN 相较于 FLTFlite 进行端侧训练时，每轮通信的平均功耗低 32%~51%。此外，对比了端侧本地训练 epoch 数量以及 batch_size 的差异对训练功耗的影响，实验结果表明增大端侧本地训练 epoch 会显著提升整体训练时长以及每轮通信的功耗。增大 batch_size 后，会降低端侧训练的时长，同时导致训练功耗上涨，FLTFlite 训练流程对端侧训练功耗增长不超过 2%，而 FedMNN 对 batch_size 的变化则比较敏感，功耗增长接近 13%。实验证明了整体使用 FedML+MNN 端云协同训练在端侧实际功耗优于 Flower+TFlite。

表 5-1 Flower+TFlite 与 FedML+MNN 训练流程端侧功耗对比

Tab.5-1 Comparison of end side power consumption between Flower+TFlite and FedML+MNN

training processes					
训练框架	epoch	batch_size	收敛时间 (mins)	每轮功耗 (mAh)	每轮功耗 增量(mAh)
FLTFlitelocal	3	100	17.82	8.8589	4.3181
FLTFlitelocal	1	100	8.99	4.5292	2.2649
FLTFlitelocal	1	1000	8.04	4.3096	2.2908
FedMNNlocal	3	100	18.08	8.6584	2.9465
FedMNNlocal	1	100	9.63	4.1511	1.1283
FedMNNlocal	1	1000	8.64	3.9922	1.2832

5.4.2 DSSM 与 Deep&Wide 模型端云协同训练

现有的 FedMNN 通常应用在图像相关任务。为了验证基于 FedMNN 的端云协同训练的有效性，以 DSSM 以及 Deep&Wide 两种推荐算法在联邦学习的设置下进行了实验。云侧服务均部署在 PC 设备上，实验过程使用 FedAvg 算法，端云通信轮数均设置为 5，端侧本地训练 epoch 均设置为 2。实验结果均使用最后一次聚合的模型进行评估。

Movielens1M 数据集数据划分按照时间以晚于 2003-01-01 的数据为测试集，早于 2003-01-01 的数据集为训练集。同时为确保测试集中的用户在训练集中有历史行为，需要过滤掉在训练集中没有行为的用户。随后，对训练数据进行随机负采样，将用户评分的电影作为正样本，在所有的电影中随机采样若干个电影作

为用户的负样本，默认正负样本比例为 1:3；随后在采样后的数据中，随机抽取互相独立的 50W 条数据作为不同端侧设备的训练数据。Avazu 数据集预处理参考 torch-rechub 中关于 Avazu 数据集的预处理。由于训练集数量过大，从训练集中随机抽取互相独立的 30W 条数据作为不同端侧设备的训练数据。从表 5-2 以及表 5-3 的结果可以看出，将端云协同训练用于推荐领域，在端侧设备持有不同数据的情况下，增加多台端侧设备进行端云协同训练，可以提高模型收敛速度，在更少的通信轮数下实现相对更高的性能。

表 5-2 Movielens1M 上 DSSM 每轮的 AUC
Tab.5-2 AUC of DSSM on Movielens1M each round

客户端数量	通信轮数	AUC
1	1	70.47
1	2	72.63
1	3	74.18
1	4	74.85
1	5	75.18
3	1	71.10
3	2	73.31
3	3	74.82
3	4	75.57
3	5	75.86

表 5-3 Avazu 上 Deep&Wide 每轮的 AUC
Tab.5-3 AUC of Deep&Wide on Avazu each round

客户端数量	通信轮数	AUC
1	1	53.45
1	2	57.08
1	3	59.27
1	4	60.68
1	5	61.67
3	1	53.97
3	2	57.70
3	3	59.85
3	4	61.19
3	5	62.11

5.6 本章小结

本章对支持隐私保护的端云协同训练框架开展研究,提出了一种融合联邦机器学习与移动神经网络的端云协同训练框架 FedMNN,本章首先介绍了 FedMNN 的整体框架与工作流程,并详细分析了云侧模型初始化、端侧序列推荐建模与本地训练、模型参数聚合与模型更新等关键环节。该框架在保障用户本地数据隐私不外泄的前提下,实现了端侧设备与云端的协同建模,有效缓解了数据孤岛问题。同时,通过对比 FedMNN 与 FLTFlite 在基准任务上不同参数设置下的训练功耗,在同等情况下,FedMNN 在每轮通信过程中的平均功耗较 FLTFlite 低 32%~51%。此外,通过在 DSSM 与 Deep&Wide 模型上进行的实验也进一步验证了 FedMNN 能够有效提升推荐准确率和模型鲁棒性,特别是在数据分布不均和设备异质性较强的场景下,依然表现出良好的适应性与稳定性。

第 6 章 总结与展望

6.1 工作总结

本文针对互联网用户行为数据具有稀疏性、异质性、动态性以及用户隐私保护需求的特点，围绕上下文增强的序列推荐算法展开研究，具体贡献主要分为以下几点：

(1) 提出了一种面向单域场景的时间上下文增强序列推荐模型 **TCSRec**。首先，针对现有序列推荐对时间信息利用不足的问题，结合位置编码和时间间隔矩阵对用户-项目交互序列进行时间感知表示增强。其次，考虑到用户兴趣随时间变化的动态性，设计了融合多头自注意力机制与时间卷积网络的多层次时序建模框架，分别从全局交互依赖和局部时间模式两个层面建模用户的动态兴趣。最后，基于注意力表示与时序特征的动态融合结果，计算下一项目的预测分布。在多个真实数据集上与主流序列推荐方法进行对比实验，结果表明 **TCSRec** 能够有效提升推荐准确率与排序性能，证明了该方法的有效性。

(2) 提出了一种面向多域场景的社交增强时间感知跨域序列推荐模型 **STRec**。首先，针对单域推荐无法全面捕捉用户跨平台兴趣迁移模式的问题，设计了跨域序列嵌入表示模块，将来自不同领域的用户交互序列进行统一嵌入建模。其次，利用图神经网络挖掘跨领域项目之间的潜在协作特征，并结合序列注意力编码器捕获跨序列的时间依赖特征。此外，考虑到社交邻居对用户兴趣演化的潜在影响，设计了社交感知注意力网络，动态融合社交邻居的序列特征，增强用户表示学习效果。在两个跨域数据集上的实验结果表明，**STRec** 能够有效缓解数据稀疏和冷启动问题，并在推荐准确率和多样性上均显著优于对比方法，验证了 **STRec** 的有效性。

(3) 提出了一种面向隐私保护场景的端云协同序列推荐框架 **FedMNN**。首先，针对传统推荐系统依赖集中数据训练带来的隐私泄露风险，设计了联邦学习与移动神经网络结合的训练机制，实现了用户隐私数据的本地化训练和联邦聚合。

其次,为兼顾推荐系统的个性化与全局泛化能力,设计了云侧模型初始化与下发、端侧序列推荐建模与本地训练、联邦参数聚合与更新等完整训练流程,确保各端设备个性化信息与全局信息的有效融合。最后,在两个真实数据集上与其他隐私保护推荐方法进行对比实验,结果表明 FedMNN 在有效保障用户隐私的同时,保持了较高的推荐准确率和系统稳定性,验证了所提方法的有效性。

6.2 未来展望

针对推荐系统现存的一些问题,本文虽然从时间上下文建模、跨域社交增强和隐私保护等角度对序列推荐算法进行了部分改进,但受限于研究时间和数据条件,算法性能和应用场景适应性仍有进一步提升的空间。针对未来的研究方向,本文有以下展望:

(1) 时间上下文信息的多维度扩展。本文对时间上下文信息的建模主要集中在时间间隔和顺序信息的挖掘,未来可以进一步考虑更丰富的时间上下文特征,例如节假日、特殊事件、用户活跃周期等时间因素,以更全面地挖掘时间信息对用户兴趣迁移的潜在影响。

(2) 跨域用户兴趣的精细化迁移。本文提出的跨域序列推荐方法主要依赖用户交互序列和社交邻居信息实现跨域用户兴趣建模,未来可以进一步探索用户在不同领域中兴趣偏好的异同特征,结合迁移学习、对比学习等技术,实现更加细粒度的用户兴趣跨域迁移与对齐。

(3) 增强隐私保护机制。本文采用联邦学习框架实现端云协同训练,虽然保障了用户行为数据不直接暴露,但仍面临通信开销、异构设备性能差异等问题。未来可以结合差分隐私、联邦蒸馏等技术,进一步提高联邦训练过程中模型的鲁棒性和隐私保护强度。

(4) 上下文信息的多源融合。本文分别在单域和跨域场景下引入了时间和社交上下文信息,未来可以进一步探索更丰富的上下文信息(如用户地理位置、使用设备、浏览路径等),构建多模态上下文增强的序列推荐模型,提升推荐系统对用户兴趣的多维度理解与动态捕捉能力。

参考文献

- [1] 第 55 次《中国互联网络发展状况统计报告》发布[J]. 传媒论坛, 2025, 9(02): 121.
- [2] Wu C, Wu F, Huang Y, et al. Personalized news recommendation: Methods and challenges[J]. ACM Transactions on Information Systems, 2023, 41(1): 1-50.
- [3] Shambour Q. A deep learning based algorithm for multi-criteria recommender systems[J]. Knowledge-based systems, 2021, 211: 106545.
- [4] LIU J, SHI C, YANG C, et al. A survey on heterogeneous information network based recommender systems: Concepts, methods, applications and resources [J]. AI Open, 2022, 3: 40-57.
- [5] 歹杰, 李青山, 褚华, 等. 突破智慧教育: 基于图学习的课程推荐系统[J]. 软件学报, 2022, 33(10): 3656-3672.
- [6] Ko H, Lee S, Park Y, et al. A survey of recommendation systems: recommendation models, techniques, and application fields[J]. Electronics, 2022, 11(1): 141.
- [7] Roy D, Dutta M. A systematic review and research perspective on recommender systems[J]. Journal of Big Data, 2022, 9(1): 1-36.
- [8] 程章桃, 钟婷, 张晟铭等. 基于图学习的推荐系统研究综述[J]. 计算机科学, 2022, 49(09): 1-13.
- [9] Wang S, Hu L, Wang Y, et al. Sequential recommender systems: challenges, progress and prospects[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 2019: 6332-6338.
- [10] Gao H, Wu Y, Xu Y, et al. Neural collaborative learning for user preference discovery from biased behavior sequences[J]. IEEE Transactions on Computational Social Systems, 2023, 11(4): 5068-5078.
- [11] Chen R, Fan J, Wu M. MC-RGN: Residual Graph Neural Networks based on Markov Chain for sequential recommendation[J]. Information Processing & Management, 2023, 60(6): 103519.
- [12] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback[C]//Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. Montreal, QC, Canada. 2009: 452-461.
- [13] Shani G, Heckerman D, Brafman R I, et al. An MDP-based recommender system[J]. Journal of Machine Learning Research, 2005, 6: 1265--1295.
- [14] Zhang Z, Nasraoui O. Efficient Hybrid Web Recommendations Based on Markov Clickstream Models and Implicit Search[C]// Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'07), CA, USA, 2007: 621-627.
- [15] Ruining He, Wang-Cheng Kang, Julian McAuley. Translation-based Recommendation[C]//Proceedings of the Eleventh Acn Conference on

- Recommender Systems, Como, Italy, 2017:161-169.
- [16] Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized markov chains for next-basket recommendation[C]//Proceedings of the 19th international conference on World wide web, North Carolina, USA, 2010: 811-820.
 - [17] Kabbur S, Ning X, Karypis G. Fism: factored item similarity models for top-n recommender systems[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago Illinois, USA, 2013: 659-667.
 - [18] Tang G, Gan X, Wang J, et al. EditKG: Editing knowledge graph for recommendation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC, USA, 2024: 112-122.
 - [19] Chen Q, Li G, Zhou Q, et al. Double attention convolutional neural network for sequential recommendation[J]. ACM Transactions on the Web, 2022, 16(4): 1-23.
 - [20] Chang J, Gao C, Zheng Y, et al. Sequential recommendation with graph neural networks[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, Online, Canada, 2021: 378-387.
 - [21] Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[EB/OL]. (2015-11-21)[2025-03-07]. <https://arxiv.org/abs/1511.06939>, 2015.
 - [22] Hidasi B, Karatzoglou A. Recurrent neural networks with top-k gains for session-based recommendations[C]//Proceedings of the 27th ACM international conference on information and knowledge management, Torino, Italy, 2018: 843-852.
 - [23] Tsai Y C, Li C T. CARE: learning convolutional attentional recurrent embedding for sequential recommendation[C]//Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Online, Netherlands, 2021: 654-660.
 - [24] Huang C, Chen J, Xia L, et al. Graph-enhanced multi-task learning of multi-level transition dynamics for session-based recommendation[C]//Proceedings of the AAAI conference on artificial intelligence, Washington DC, USA, 2021, 35(5): 4123-4130.
 - [25] Wang Z, Wei W, Cong G, et al. Global context enhanced graph neural networks for session-based recommendation[C]//Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, Online, China, 2020: 169-178.
 - [26] Tang J, Wang K. Personalized top-n sequential recommendation via convolutional sequence embedding[C]//Proceedings of the eleventh ACM

- international conference on web search and data mining, New York, USA: ACM, 2018: 565-573.
- [27] 王鸿伟, 过敏意. 刻画长短期用户兴趣的基于会话的推荐系统[J]. 中国科学: 信息科学, 2020, 50(12): 1867-1881.
- [28] 冯永, 张备, 强保华, 等. MN-HDRM: 长短兴趣多神经网络混合动态推荐模型[J]. 计算机学报, 2019, 42(01): 16-28.
- [29] Luo Y, Liu Q, Liu Z. Stan: Spatio-temporal attention network for next location recommendation[C]//Proceedings of the web conference 2021, Ljubljana Slovenia, 2021: 2177-2185.
- [30] Ma J, Zhou C, Yang H, et al. Disentangled self-supervision in sequential recommenders[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Online, USA, 2020: 483-491.
- [31] Kang K, Park J, Kim W, et al. Recommender system using sequential and global preference via attention mechanism and topic modeling[C]//Proceedings of the 28th ACM international conference on information and knowledge management, Beijing, China, 2019: 1543-1552.
- [32] Li Y, Chen T, Zhang P F, et al. Lightweight self-attentive sequential recommendation[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Queensland, Australia, 2021: 967-977.
- [33] Sun F, Liu J, Wu J, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM international conference on information and knowledge management, Beijing, China, 2019: 1441-1450.
- [34] Wu C Y, Ahmed A, Beutel A, et al. Recurrent recommender networks[C]//Proceedings of the tenth ACM international conference on web search and data mining. Cambridge, United Kingdom, 2017: 495-503.
- [35] Liu Q, Zeng Y, Mokhosi R, et al. STAMP: short-term attention/memory priority model for session-based recommendation[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, London, United Kingdom, 2018: 1831-1839.
- [36] 袁涛, 牛树梓, 李会元. 一种基于 CW-RNN 的多时间尺度序列建模推荐算法[J]. 中文信息学报, 2020, 34(06): 97-105.
- [37] Li J, Wang Y, McAuley J. Time interval aware self-attention for sequential recommendation[C]//Proceedings of the 13th international conference on web search and data mining, Houston TX, USA, 2020: 322-330.
- [38] Ji W, Wang K, Wang X, et al. Sequential recommender via time-aware attentive memory network[C]//Proceedings of the 29th ACM international conference on information & knowledge management, Online, Ireland, 2020: 565-574.
- [39] Xia L, Huang C, Xu Y, et al. Multi-behavior sequential recommendation with temporal graph transformer[J]. IEEE Transactions on Knowledge and Data

- Engineering, 2022, 35(6): 6099-6112.
- [40] Yang Y, Huang C, Xia L, et al. Multi-behavior hypergraph-enhanced transformer for sequential recommendation[C]//Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining, Washington DC, USA, 2022: 2263-2274.
- [41] 陈毓哲,曹琼,黄贤英,等.MB-ATMK: 融合属性权重和时序元知识的多行为序列推荐模型[J].计算机科学,2024,51(S2):617-625.
- [42] Su J, Chen C, Lin Z, et al. Personalized behavior-aware transformer for multi-behavior sequential recommendation[C]//Proceedings of the 31st ACM international conference on multimedia, Ottawa ON, Canada, 2023: 6321-6331.
- [43] Cao J, Cong X, Liu T, et al. Item similarity mining for multi-market recommendation[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 2022: 2249-2254.
- [44] Zhang W, Zhang P, Zhang B, et al. A Collaborative Transfer Learning Framework for Cross-domain Recommendation[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach CA, USA, 2023: 5576-5585.
- [45] Liu M, Li J, Li G, et al. Cross domain recommendation via bi-directional transfer graph collaborative filtering networks[C]//Proceedings of the 29th ACM international conference on information & knowledge management, Ireland, 2020: 885-894.
- [46] Ma M, Ren P, Lin Y, et al. π -net: A parallel information-sharing network for shared-account cross-domain sequential recommendations[C]//Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, Paris, France, 2019: 685-694.
- [47] Sun W, Ma M, Ren P, et al. Parallel split-join networks for shared account cross-domain sequential recommendations[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(4): 4106-4123.
- [48] Zhuang F Z, Zhou Y M, Ying H C, et al. Sequential recommendation via cross-domain novelty seeking trait mining[J]. Journal of Computer Science and Technology, 2020, 35: 305-319.
- [49] Guo L, Tang L, Chen T, et al. DA-GCN: a domain-aware attentive graph convolution network for shared-account cross-domain sequential recommendation[C]//Proceedings of the 30th International Joint Conference on Artificial Intelligence, Montreal, 2021: 2483-2489.
- [50] Ma M, Ren P, Chen Z, et al. Mixed information flow for cross-domain sequential recommendations[J]. ACM Transactions on Knowledge Discovery from Data, 2022, 16(4): 1-32.
- [51] Qiao P, Zhang Z, Li Z, et al. TAG: Joint triple-hierarchical attention and GCN for review-based social recommender system[J]. IEEE Transactions on

- Knowledge and Data Engineering, 2022, 35(10): 9904-9919.
- [52] Oubalahcen H, El Ouadghiri M D. Recommender Systems for Social Networks: A Short Review[C]//Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security, Larache, Morocco, 2023: 1-7.
- [53] Narayanan P, Vivekanandan K. Hybrid CNN and RNN-based shilling attack framework in social recommender networks[J]. EAI Endorsed Transactions on Scalable Information Systems, 2022, 9(35).
- [54] 张青博,王斌,崔宁宁,等.基于注意力机制的规范化矩阵分解推荐算法[J].软件学报,2020,31(03):778-793.
- [55] Ma H, Yang H, Lyu M R, et al. Sorec: social recommendation using probabilistic matrix factorization[C]//Proceedings of the 17th ACM conference on Information and knowledge management, California, USA, 2008: 931-940.
- [56] Yang B, Lei Y, Liu J, et al. Social collaborative filtering by trust[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(8): 1633-1647.
- [57] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks[C]//Proceedings of the fourth ACM conference on Recommender systems, Barcelona, Spain, 2010: 135-142.
- [58] Pan Y, He F, Yu H. Learning social representations with deep autoencoder for recommender system[J]. World Wide Web, 2020, 23(4): 2259-2279.
- [59] 温涛,张敏,王怀远.基于堆叠稀疏降噪自编码器的暂态稳定评估模型[J].电力工程技术,2022,41(01):207-212.
- [60] Wan L, Xia F, Kong X, et al. Deep matrix factorization for trust-aware recommendation in social networks[J]. IEEE Transactions on Network Science and Engineering, 2020, 8(1): 511-528.
- [61] Wu B, Zhong L, Yao L, et al. EAGCN: An efficient adaptive graph convolutional network for item recommendation in social Internet of Things[J]. IEEE internet of things journal, 2022, 9(17): 16386-16401.
- [62] Song W, Xiao Z, Wang Y, et al. Session-based social recommendation via dynamic graph attention networks[C]//Proceedings of the Twelfth ACM international conference on web search and data mining, Melbourne VIC, Australia, 2019: 555-563.
- [63] Zheng X, Li X, Chen Z, et al. Enhanced self-attention mechanism for long and short term sequential recommendation models[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(3): 2457-2466.
- [64] Zhaoyan H, Yonglong L, Xiaoyao Z, et al. A novel privacy-preserving matrix factorization recommendation system based on random perturbation[J]. Journal of Intelligent & Fuzzy Systems, 2020, 38(4): 4525-4535.
- [65] Zhou H, Yang G, Xiang Y, et al. A lightweight matrix factorization for recommendation with local differential privacy in big data[J]. IEEE Transactions on Big Data, 2021, 9(1): 160-173.

- [66] Zhang L, Zhu X, Ma J, et al. Medical privacy-preserving service recommendation[C]// Proceedings of the 20th IEEE International Conference on Communication, Dublin, Ireland, 2020: 1-6.
- [67] Zhang C, Xie Y, Bai H, et al. A survey on federated learning[J]. Knowledge-Based Systems, 2021, 216: 106775.
- [68] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design[J]. Machine Learning and Systems, 2019, 1: 374-388.
- [69] Liu Y, Fan T, Chen T, et al. FATE: An industrial grade platform for collaborative learning with data protection[J]. The Journal of Machine Learning Research, 2021, 22(1): 10320-10325.
- [70] He C Y, Li S Z, So J, et al. FedML: A research library and benchmark for federated machine learning [EB/OL]. (2020-11-08)[2025-03-07]. <https://arxiv.org/abs/2007.13518>, 2020.
- [71] Li K H, de Gusmão P P B, Beutel D J, et al. Secure aggregation for federated learning in flower[C]//Proceedings of the 2nd ACM International Workshop on Distributed Machine Learning, Germany, 2021: 8-14.
- [72] 林伟伟,石方,曾岚等.联邦学习开源框架综述[J].计算机研究与发展,2023,60(07):1551-1580.
- [73] Wang L, Wang Y, Bai Y, et al. POI recommendation with federated learning and privacy preserving in cross domain recommendation[C]//Proceedings of the IEEE Conference on Computer Communications Workshops, Vancouver, BC, Canada, 2021: 1-6.
- [74] Zhao Q, Zhao C, Cui S, et al. PrivateDL: privacy-preserving collaborative deep learning against leakage from gradient sharing[J]. International Journal of Intelligent Systems, 2020, 35(8): 1262-1279.
- [75] Wei K, Li J, Ding M, et al. Federated learning with differential privacy: Algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454-3469.
- [76] Fang H, Qian Q. Privacy preserving machine learning with homomorphic encryption and federated learning[J]. Future Internet, 2021, 13(4): 94.
- [77] Truex S, Baracaldo N, Anwar A, et al. A hybrid approach to privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, London, United Kingdom, 2019. 1-11.
- [78] Ni S, Hu S, Li L. An intention-aware markov chain based method for top-k recommendation[J]. IEEE Transactions on Automation Science and Engineering, 2022, 21(1): 581-592.
- [79] Cui Q, Wu S, Liu Q, et al. MV-RNN: A multi-view recurrent neural network for sequential recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 32(2): 317-331.
- [80] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

- [81] Zivic P, Vazquez H, Sánchez J. Scaling Sequential Recommendation Models with Transformers[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC, USA, 2024: 1567-1577.
- [82] Kang W C, McAuley J. Self-attentive sequential recommendation[C]//Proceedings of the 18th IEEE International Conference on Data Mining (ICDM), Singapore, 2018: 197-206.
- [83] Yu L, Zhang C, Liang S, et al. Multi-order attentive ranking model for sequential recommendation[C]//Proceedings of the AAAI conference on artificial intelligence, Hawaii, USA, 2019: 5709-5716.
- [84] Wang T, Wang C, Tian H, et al. GeoMixer: The MLP-Based Sequential POI Recommender with Travel Routing Modelling[C]//Proceedings of the 2023 IEEE International Conference on Data Mining (ICDM), Shanghai, China, 2023: 1373-1378.
- [85] Nasir M, Ezeife C I. A survey and taxonomy of sequential recommender systems for e-commerce product recommendation[J]. SN Computer Science, 2023, 4(6): 708.
- [86] Zhang Q, Wu B, Sun Z, et al. Gating augmented capsule network for sequential recommendation[J]. Knowledge-Based Systems, 2022, 247: 108817.
- [87] Yuan E, Guo W, He Z, et al. Multi-behavior sequential transformer recommender[C]//Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, Madrid, Spain, 2022: 1642-1652.
- [88] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web, Hong Kong, China, 2001: 285-295.
- [89] Haffar R, Sanchez D, Domingo-Ferrer J. Explaining predictions and attacks in federated learning via random forests[J]. Applied Intelligence, 2023, 53(1): 169-185.
- [90] Zhao B, Mopuri K R, Bilen H. idlg: Improved deep leakage from gradients[EB/OL]. (2020-01-08)[2025-03-07]. <https://arxiv.org/abs/2001.02610>, 2020.
- [91] Li T, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE signal processing magazine, 2020, 37(3): 50-60.
- [92] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundations and trends® in machine learning, 2021, 14(1-2): 1-210.
- [93] Mathur A, Beutel D J, De Gusmão P P B, et al. On-device federated learning with flower[EB/OL]. (2021-04-07)[2025-03-07]. <https://arxiv.org/abs/2104.03042>, 2021.
- [94] Jiang X T, Wang H, Chen Y L, et al. MNN: A universal and efficient

inference engine[EB/OL]. (2020-02-27)[2025-03-07]. <https://arxiv.org/abs/2002.12418>, 2020.

攻读学位期间发表的学术论文目录

已录用的论文:

- [1] 高祥云, 孟丹, 罗明凯, 王俊, 张丽平, 孔超.支持隐私保护的端云协同训练[J].华东师范大学学报(自然科学版),2023,(05):77-89. (第一作者, 北大核心、CSCD)
- [2] Gao X, Meng D, Sun X, et al. STRec: Social-Augmented Time-Aware Cross-Domain Sequential Recommendation[C]//2024 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA). IEEE, 2024: 502-509. (第一作者, CCF C, EI)
- [3] 孔超, 高祥云, 孙先兰, 陈家会, 张丽平. 融合学习者遗忘行为的多特征深度知识追踪[J].桂林电子科技大学学报. 2024 (第二作者, 已录用)

攻读学位期间参与的科研项目

- [1] 国家自然科学基金青年科学基金项目, 面向互联网开放数据的实体匹配及其应用研究 (项目号 61902001)
- [2] 安徽省高等学校科学研究项目 (自然科学类), 面向数据安全性与隐私保护的个性化推荐方法研究 (项目号 2023AH050914)

致 谢

行文至此，意味着我的研究生求学阶段即将画上句点。回首过往三年，科研路上的迷茫与顿悟、挫败与欣喜，此刻皆化作沉淀于心的感动与感恩。在此，谨以最诚挚的谢意向所有给予我支持与帮助的师长、同窗及亲友致以敬意。

首先，我要衷心感谢我的导师孔超老师。在我攻读硕士学位的三年里，孔超老师以其渊博的学识、严谨的治学态度以及宽广的学术视野，给予我无微不至的关怀和悉心指导。在我初入研究领域时，孔超老师耐心引导我熟悉研究方法，帮助我明确研究方向。并在论文的选题、模型设计及论文撰写过程中都提出了宝贵的意见。在孔超老师的指导下，我最终顺利完成了本课题的研究工作。孔超老师的谆谆教诲将指引我在今后的学术和工作道路上继续前行。还要衷心感谢计算机与信息学院的各位老师的悉心教导，各位老师们在课程学习与实验开展中给予的专业指导，老师们严谨治学的态度和诲人不倦的精神，让我在专业知识与技能上取得了长足进步。

同时，我也要感谢课题组的李邵莹师姐、孙先兰、陈家会等伙伴，正是大家在科研中的相互交流和探讨，让我在研究过程中少走了许多弯路，在我遇到技术难题或论文写作困惑时，大家总是毫无保留地分享经验，给予我鼓励和帮助。此外，非常感谢 720 实验室的王潇冉、金晶、彭杰、董纬和祝玉宣等伙伴，感谢你们让实验室充满活力与欢笑。无论是日常的学术讨论，还是闲暇时的交流谈心，都让我感受到这个集体的温暖。你们的勤奋与热情也时常激励着我不断前进。

在此，我还要感谢我的家人们。在我求学的三年里，家人们始终默默支持着我，让我无后顾之忧地专注于学业。每当我遇到挫折或感到疲惫时，家人的关心和鼓励给予了我无限的力量和动力，让我坚持不懈地走到今天。

特别感谢参与论文评审与答辩的各位专家，感谢您们在百忙中拨冗指导。文虽终而学无涯，未来我将带着这段珍贵岁月赋予我的知识与信念，继续在其他领域探索 and 追求，将所学应用于实践，为社会贡献自己的力量。

再次向所有曾帮助、支持和鼓励过我的人们，致以最诚挚的感谢！