计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(7): 1611−1621, 2025

基于大语言模型的可信多模态推荐算法

闫 萌 徐 偲 黄海槟 赵 伟 管子玉
(西安电子科技大学计算机科学与技术学院 西安 710126)
(mengyan@stu.xidian.edu)

Large Language Model-Based Trusted Multi-Modal Recommendation Algorithm

Yan Meng, Xu Cai, Huang Haibin, Zhao Wei, and Guan Ziyu
(*School of Computer Science and Technology*, *Xidian University*, *Xi'an* 710126)

Abstract Sequential recommendation is centered on mining users' preferences and behavior patterns from their interaction sequences. Existing work has recognized the inadequacy of single-modal interaction data, and has utilized a large amount of multi-modal data, including item reviews, homepage images, and other sources, to complement interaction data and improve recommendation performance. However, these multi-modal data are often interspersed with unavoidable noise that may limit the exploration of personalized user preferences. While suppressing inter-modal inconsistent information can reduce noise interference, it is almost impossible to completely eliminate noise from user-generated multi-modal content. To address the above challenges, we propose a large language model-based trusted multi-modal recommendation (Large-TR) algorithm, which aims to provide the trustworthy recommendation in noisy multi-modal data scenarios. Specifically, the algorithm relies on the excellent natural language understanding capability of the large language model, which efficiently filters the noise in multi-modal data and achieves more accurate and detailed modelling of user preferences. Additionally, we design a trustworthy decision mechanism that dynamically evaluates the uncertainty of recommendation results and ensures the usability of recommendation results in high-risk scenarios. Experimental results on four widely used public datasets show that the algorithm proposed in this paper has better performance compared with other baseline algorithms. Our source code is available at https://github.com/ hhbray/Large-TR.

Key words sequential recommendation; multi-modal; user-generated content; trustworthy decision; large language model

摘 要 序列推荐的核心在于从用户的交互序列中挖掘其偏好和行为模式.现有研究已经认识到单一模态交互数据存在不足,因此借助大量多模态数据(如商品评价、主页图片等)来丰富交互信息,提升推荐系统的性能.然而,这些多模态数据中常常夹杂着不可避免的噪声,可能会限制用户个性化偏好的探索.尽管可以通过抑制模态间不一致的信息来减少噪声干扰,但要完全消除用户生成的多模态内容中的噪声几乎是不可能的.针对上述挑战,提出了一种基于大语言模型的可信多模态推荐算法,旨在于含噪多模态数

收稿日期: 2024-06-03; **修回日期**: 2024-11-21
基金项目: 国家自然科学基金项目(62133012, 61936006, 62103314, 62073255, 62303366); 陕西省重点研发计划项目(2020ZDLGY04-07); 陕西省创新能力支撑计划项目(2021TD-05); 陕西省自然科学基础研究计划项目(2023-JC-QN-0648); 中央高校基本科研业务费专项资金: 西安电子科技大学研究生创新基金(YJSJ25012)
This work was supported by the National Natural Science Foundation of China (62133012, 61936006, 62103314, 62073255, 62303366), the Key Research and Development Program of Shaanxi Province (2020ZDLGY04-07), the Innovation Capability Support Program of Shaanxi Province (2021TD-05), the Natural Science Basic Research Program of Shaanxi Province (2023-JC-QN-0648), and the Fundamental Research Funds for the Central Universities: the Innovation Fund of Xidian University (YJSJ25012).
通信作者: 徐偲(cxu@xidian.edu.cn)

据场景下提供可信的推荐结果. 具体而言, 该算法依托于大语言模型卓越的自然语言理解能力, 高效过滤多模态数据中的噪声, 实现对用户偏好更为精确和细致的建模. 此外, 还设计了一种可信决策机制, 用于动态评估推荐结果的不确定性, 以确保在高风险场景下推荐结果的可用性. 在 4 个广泛使用的公开数据集上的实验结果显示, 相较于其他基线算法, 提出的算法有更好的性能表现. 代码可以在 https://github.com/hhbray/Large-TR 获取.

关键词 序列推荐; 多模态; 用户生成内容; 可信决策; 大语言模型

中图法分类号 TP37

DOI: 10.7544/issn1000-1239.202440433 **CSTR:** 32373.14.issn1000-1239.202440433

随着互联网服务平台进入存量增长阶段, 已注册用户不断产生各种交互行为. 在这种情况下, 为现有用户提供更精准的服务已经成为大多数在线服务平台的首要任务. 序列推荐可以显式地建模用户的序列行为, 对未来时刻用户可能感兴趣的物品进行预测. 早期的研究[1]根据购买相似物品的用户群来建立相关性模型. 然而, 用户交互的项目数量与项目总数相比微不足道, 导致严重的稀疏性问题. 许多研究[2-4]引入项目侧信息(如类别、标题、主页图像等)来丰富项目表示. 尽管可以在一定程度上缓解稀疏性问题, 然而这类信息都是商家按照模板上传生成的, 缺乏多样性.

互联网上存在着大量丰富的用户生成内容, 这是一种用户自发创造的多模态信息, 包括评论、图片、视频、评级等, 可以作为购买和浏览等基本交互的补充. 研究人员尝试从用户生成内容中挖掘用户偏好, 以应对稀疏性问题的挑战[5-8], 然而, 数据中存在的广告或误操作等噪声限制了推荐系统的性能. 如图 1 中, 用户评论中存在着“绿色”“优惠”等与真实偏好无关的噪声, 影响用户文本偏好的捕获. 部分工作[9-10]利用不同模态间的一致性信息来减少噪声的干扰, 但必须承认的是, 完全消除用户生成内容中的噪声是不现实的.



Fig. 1 Examples of noise in user-generated content

图 1 用户生成内容中噪声的示例

大语言模型(large language model, LLM)的兴起为推荐系统注入了新的活力, 也为解决噪声问题带来了新的可能. 大语言模型可以进行深度的语义分析, 理解用户的隐含意图, 生成个性化的推荐. 但大语言模型的训练数据与推荐场景并不相关, 语义空间难对齐. 部分工作试图通过调整指令来引导大语言模型生成更准确的推荐结果, 但结果容易被干扰, 还可能会出现幻觉问题.

针对上述问题, 本文提出了一种基于大语言模型的可信多模态推荐(large language model-based trusted multi-modal recommendation, Large-TR)算法, 它利用大语言模型强大的语义理解能力过滤多模态数据中的噪声, 对多模态用户偏好进行建模. 此外, 针对噪声无法完全消除的问题, 我们还设计了一种可信决策机制, 动态评估推荐结果的不确定性, 在高风险场景下保证推荐结果的可用性. 在 4 个广泛使用的公开数据集上的实验结果显示, 相对于其他基线算法, 本文提出的算法有更好的性能表现.

综上, 本文的主要贡献包括 3 个方面:

1)借助大语言模型过滤低质量多模态信息中的噪声, 建模更精确和细致的用户偏好, 有效提高推荐性能;

2)设计了一种简单但有效的可信决策机制, 可以动态评估推荐结果的可信度, 保证推荐系统在有噪声的情况下依然可用;

3) 在 4 个广泛使用的真实数据集上进行了实验, 实验结果表明本文提出的算法可以显著提高性能.

1 相关工作

1.1 多模态序列推荐

序列推荐是从用户的历史交互序列中建模用户

偏好的演变过程，预测下一个可能交互的商品. 早期序列推荐的工作[11-14]大多基于马尔可夫链假设，但稀疏性问题严重限制了性能的上限. Zhou 等人[15]尝试建模项目与属性之间的关系，以增强项目的表示. Xie 等人[16]从商品的描述、标题等文本信息中挖掘用户偏好，提升推荐性能. 很多情况下，图像作为一种视觉内容，可以提供语言无法表达的视觉信息. 受此启发，有工作[17]提出将视觉信号纳入用户偏好的建模过程，预测下一个可能交互的商品. 近年来，微视频作为电商平台新兴的内容形式，包含多场景、多角度的商品展示，Lei 等人[18]提出了基于微视频的序列推荐算法. 综上，基于多模态的序列推荐已经开始集成来自图像、文本和其他模态（如视频和音频）的信息，以实现对项目更全面的表示和对用户偏好更准确的理解. 但现有方法仍面临 2 个主要问题：1）仅依赖商品相关的侧信息来丰富商品表征，未能直接对用户偏好进行建模，导致对用户需求的理解不够深入；2）多模态信息中往往存在不匹配不规范的噪声数据，严重影响推荐系统的性能. 因此本文提出了一种基于用户生成内容的多模态序列推荐方法，通过大语言模型丰富的语义理解能力对多模态内容降噪，从用户的角度挖掘序列偏好和行为趋势.

1.2 基于大语言模型的推荐

语言模型在推荐系统的发展中扮演着至关重要的角色. 早期研究[19-20]已经发现，语言模型框架具有统一建模各类推荐任务的潜力. 随着大语言模型的出现，推荐领域迎来了巨大的进步. 这些大语言模型具备极强的语言理解、生成和泛化能力，为解决推荐系统中的冷启动、个性化推荐和可解释性等问题提供了新的思路和方法. 但实现这一目标通常需要大量的推荐数据进行训练. 2023 年，Dai 等人[21]通过特定的指令激发大语言模型直接生成推荐结果，进一步证实了大语言模型在推荐领域的巨大潜力. 在此基础上，部分研究[22-23]对指令进行微调，以期望将大语言模型与推荐系统的语义空间尽可能对齐，从而获得更好的推荐效果. 基于大语言模型的推荐方法仍处于探索阶段，其性能很大程度上依赖于指令的质量，生成结果易受干扰，如何确保生成鲁棒的推荐结果是未来需要重点关注的问题. 因此，本文并没有通过大语言模型直接生成推荐结果，而是利用大语言模型已表现出的强大的语义理解与分析能力，过滤多模态交互数据中存在的噪声，将多模态信息和大语言模型的知识共同注入到推荐系统中，提高推荐结果的准确性和鲁棒性.

2 问题定义

给定用户 U 的历史交互序列 $S^U=\{I_1,I_2,\cdots,I_N\}$，用户 U 对序列中每个商品的文本评论可以表示为 $R^U=\{r_1^U,r_2^U,\cdots,r_N^U\}$，相应地，商品 I 收到的所有文本评论可以表示为 $R^I=\{r_1^I,r_2^I,\cdots,r_M^I\}$. 本文收集交互序列中每个商品的首页图片作为多模态内容中的图片信息，表示为 $G^U=\{g_1^U,g_2^U,\cdots,g_N^U\}$. 因此，本文的目标是通过观察到的交互序列 S^U 和多模态内容 (R^U,R^I,G^U)，预测出用户与未观察到的商品交互的概率，并提供预测结果的不确定性.

3 基于大语言模型的可信多模态推荐算法

本文提出的算法模型整体架构 Large-TR 如图 2 所示，包括大语言模型去噪模块、多视图序列偏好模块和可信决策模块 3 个部分. Large-TR 的设计基于 3 个假设：1）大语言模型有理解图像和文本语义的能力；2）用户生成的多模态信息是揭示用户偏好的重要补充；3）用户侧和商品侧都可以获得有效的推荐结果.

3.1 大语言模型去噪

用户的文本评论是体现个人偏好的重要信息来源，但其中存在的广告、误评等噪声屡见不鲜. 商品图片作为商家上传的模板化内容，其内容质量更高，还可以提供语言无法表达的视觉信息. 分析图像和文本之间的关联性，我们可以更好地判断评论的真实性和准确性. 大语言模型在图文匹配的稀疏场景下具有很强的泛化能力，并在少量甚至零样本任务中实现了非常可观的性能. 我们假设大语言模型已经具有理解图文信息、判断其是否一致的能力，我们需要做的就是通过调整提示符来触发这种能力. 提示符是提供给大语言模型的文本片段，用来设定上下文或指示大语言模型如何继续生成文本，广泛应用于各种自然语言处理任务，如文本生成、文档总结等. 本文旨在利用大语言模型强大的图文理解能力，判断用户的评论文本与商品图片是否相关，从而过滤文本中的噪声. 为此，我们参考了相关上下文学习[24]和指令调整[22]的工作，并将这种能力表示为具有特定领域提示的任务.

图 2 左侧说明了我们如何调整提示符从大语言模型中引出图文匹配的功能，提示由 2 个部分组成：

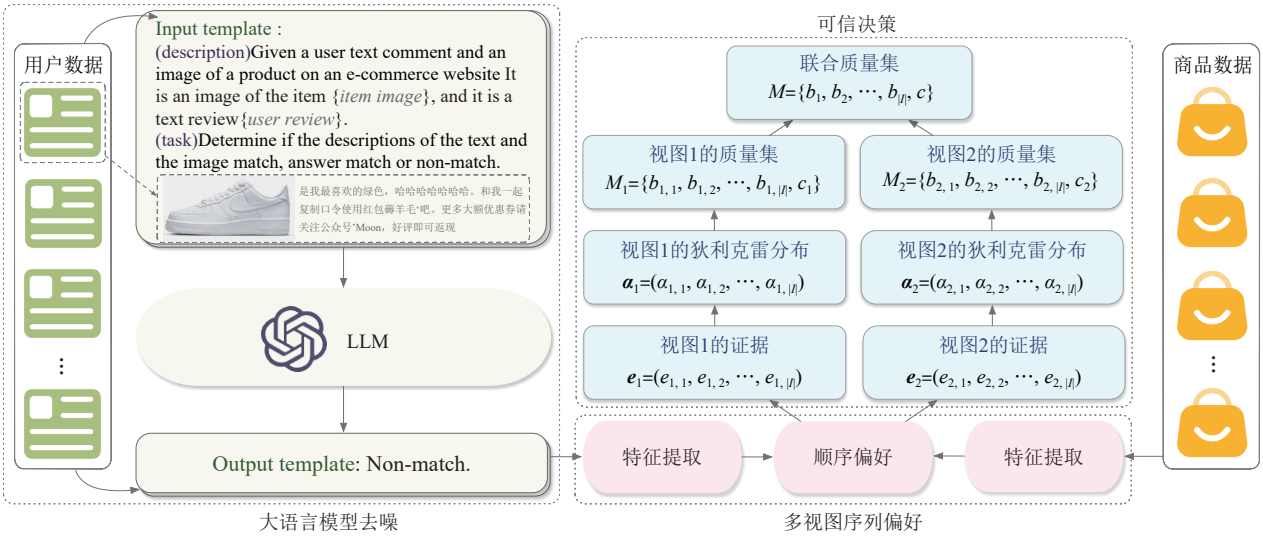


Fig. 2 Structure diagram of Large-TR

图 2 Large-TR 的结构图

1)描述. 需要设计领域相关的文本，增强大语言模型对执行相关任务所需知识的感知能力. 2)任务. 给定用户的评论与相应的商品图片，要求大语言模型给出回答. 在优化过程中，我们会为大语言模型添加示例，以帮助其准确理解任务要求. 随后，通过真实标签对模型输出进行评估，并基于评估结果不断改进提示符的设计，以提升模型的图文匹配能力. 相关的提示符优化过程在附录 A 中给出.

为了更有效地将大语言模型过滤后的多模态信息用于后续的推荐任务，我们利用预训练模型对文本和图片进行嵌入处理. 对于文本，我们将输入的文本句子分割成多个词汇单元(*token*)，并在句子的开头添加 [CLS] 和 [SEP] 等特殊标记. 预训练模型会输出 [CLS] 的隐藏状态，它可以表示整个句子的语义信息，用于下游任务. 具体来说，我们将 r_n^U 和 r_m^I 视为 *token* 序列，即 $r_n^U=\{token_k\}_{k=1}^{|r_n^U|}$，$r_m^I=\{token_k\}_{k=1}^{|r_m^I|}$，每个评论按如下方式进行嵌入：

$$\boldsymbol{t}_n^U=F_{\mathrm{t}}\left(r_n^U\right),\boldsymbol{t}_m^I=F_{\mathrm{t}}\left(r_m^I\right),\boldsymbol{t}_n^U,\boldsymbol{t}_m^I\in\mathbb{R}^{d_{\mathrm{t}}},\tag{1}$$

$F_{\mathrm{t}}(\cdot)$是 BERT(bidirectional encoder representations from Transformers)模型[25]. 根据式(1)可以获得文本嵌入 $\boldsymbol{T}^U=(\boldsymbol{t}_1^U,\boldsymbol{t}_2^U,\cdots,\boldsymbol{t}_N^U)$和 $\boldsymbol{T}^I=(\boldsymbol{t}_1^I,\boldsymbol{t}_2^I,\cdots,\boldsymbol{t}_M^I)$. 通过汇总和分析不同用户对同一商品的多样化评论，我们可以提取到更全面、更丰富的商品特征信息，还可以在一定程度上过滤掉那些与商品无关的用户文本偏好. 基于上述考虑，我们通过共同注意力机制[26](co-attention mechanism)提取 $\boldsymbol{R}^U$ 和 $\boldsymbol{R}^I$ 之间的匹配模式，对用户的文本偏好进行重新建模. 具体来说，该机制分为 3 个步骤：

1) 构建亲和力矩阵

$$\boldsymbol{A}^{UI}=\tanh\left(\left(\boldsymbol{T}^I\right)^{\mathrm{T}}\boldsymbol{M}^R\boldsymbol{T}^U\right),\boldsymbol{A}^{UI}\in(-1,1)^{M\times N},\tag{2}$$

其中 $\boldsymbol{M}^R$ 是一个参数矩阵，$\boldsymbol{A}^{UI}$ 中第 i 行第 j 列的元素表示 $\boldsymbol{R}^I$ 第 i 个评论与 $\boldsymbol{R}^U$ 第 j 个评论之间的相似性.

2)我们对 $\boldsymbol{A}^{UI}$ 实施行最大化操作，然后使用 *softmax* 生成相关性向量

$$\boldsymbol{a}^{UI}=softmax\left(RowMax\left(\boldsymbol{A}^{UI}\right)\right),\boldsymbol{a}^{UI}\in(0,1)^M.\tag{3}$$

3)对商品评论矩阵进行关注度聚合，计算最终的相关性嵌入，并与原文进行拼接.

$$\tilde{\boldsymbol{t}}_n^U=\boldsymbol{t}_n^U\oplus\boldsymbol{a}^{UI}\boldsymbol{T}^I,\tilde{\boldsymbol{t}}_n^U\in\mathbb{R}^{d_{\mathrm{t}}},\tag{4}$$

其中⊕代表向量连接操作，这一步代表对 S^U 中的每个 r_n^U 重新建模以获得重构的用户文本偏好：$\tilde{\boldsymbol{T}}^U=(\tilde{\boldsymbol{t}}_1^U,\tilde{\boldsymbol{t}}_2^U,\cdots,\tilde{\boldsymbol{t}}_N^U)$. 许多偏好信息通常很难通过文字准确传达，图像可以提供直观的视觉展示，捕捉到用户对物品外观、风格、颜色等方面的偏好. 与文本嵌入方式类似，我们使用预训练模型对每张图片进行嵌入：

$$\boldsymbol{z}_n^U=F_{\mathrm{v}}\left(g_n^U\right),\boldsymbol{z}_n^U\in\mathbb{R}^{d_{\mathrm{v}}},\tag{5}$$

其中 $F_{\mathrm{v}}(\cdot)$是 CLIP(contrastive language-image pre-training)模型[27]，用户视觉偏好可以表示为 $\boldsymbol{Z}^U=(\boldsymbol{z}_1^U,\boldsymbol{z}_2^U,\cdots,\boldsymbol{z}_N^U)$. 图文嵌入模块均支持灵活的嵌入维度，也支持扩展到更多的模态. 这些预训练模型可以看作是对模态信息进行特征提取的编码器. 文本偏好和视觉偏好的结合可以增强对用户多层次、多维度偏好的理解. 因此，我们将文本偏好和视觉偏好作为多模态视角进行整合：

$$\boldsymbol{x}_1^U=\boldsymbol{W}_{\mathrm{m}}\left[\boldsymbol{Z}^U\oplus\tilde{\boldsymbol{T}}^U\right],\boldsymbol{x}_1^U\in\mathbb{R}^{d_{\mathrm{m}}\times N},\tag{6}$$

其中 $\boldsymbol{W}_{\mathrm{m}}\in\mathbb{R}^{d_{\mathrm{m}}\times(d_{\mathrm{t}}+d_{\mathrm{v}})}$ 代表权重矩阵. 为了尽可能地让用

户的文字偏好和视觉偏好保持对齐，减少文本偏好可能带来的噪声干扰，本文使用跨模态损失函数[28]. 图像-文本对可以构造为$\{(r_i^U, g_j^U), o_{ij}\}_{i,j=1}^N$，其中$o_{ij}$=1表示图文是匹配的. 因此，图文匹配的投影概率为

$$p_{ij} = \frac{\exp((\tilde{\boldsymbol{t}}_i^U)^{\mathrm{T}} \bar{\boldsymbol{z}}_j^U)}{\sum_{k=1}^{N} \exp((\tilde{\boldsymbol{t}}_i^U)^{\mathrm{T}} \bar{\boldsymbol{z}}_k^U)}, \tag{7}$$

其中$\bar{\boldsymbol{z}}_j^U = \frac{\boldsymbol{z}_j^U}{\|\boldsymbol{z}_j^U\|}$代表归一化的图像表示，$(\tilde{\boldsymbol{t}}_i^U)^{\mathrm{T}} \bar{\boldsymbol{z}}_j^U$是文本偏好$\tilde{\boldsymbol{t}}_i^U$在图像偏好$\boldsymbol{z}_j^U$上的标量投影. 通过计算投影概率$p_{ij}$和真实匹配概率$q_{ij}$之间的KL散度（Kullback-Leibler），得到文本与图像间的损失函数：

$$\mathcal{L}_{\mathrm{rec}}(S^U) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} \ln \frac{p_{ij}}{q_{ij} + \varepsilon}, \tag{8}$$

其中ε是一个非常小的用于避免数值溢出的问题的数，$q_{ij} = \frac{o_{ij}}{\sum_{k=1}^{N} p_{ik}}$代表$(r_i^U, g_j^U)$归一化的真实匹配概率. 通过最小化损失函数，可以使投影概率分布曲线的形状接近真实匹配概率分布曲线的形状，从而使得匹配的图像文本对的投影值最大，不匹配的图像文本对的投影值最小.

3.2 多视图序列偏好

在推荐领域，使用项目唯一标识符（identity, ID）对项目进行表示一直是主流做法. 本文认可项目ID的作用，也希望尽可能地保留原始项目ID的信息不受其他因素的干扰. 因此，本文将其视为一个独立的视图. 具体来说，将交互序列S^U输入到项目嵌入层，以获得商品视角的嵌入：

$$\boldsymbol{x}_2^U = F_i(S^U), \boldsymbol{x}_2^U \in \mathbb{R}^{d_i \times N}. \tag{9}$$

众所周知，项目ID和多模态信息都能提供有效的推荐，然而，仅仅将两者拼接或者融合在一起会丢失有价值的语义信息. 因此我们认为应该从互补但独立的视角$\boldsymbol{x}_v^U$来考虑最终的推荐结果，$v=1$代表多模态信息的视角，$v=2$代表项目ID的视角，视角的数量可以随着推荐系统的需求进一步扩展. 我们使用门控循环单元（gated recurrent unit, *GRU*）来捕获各个视图中的顺序信息：

$$\boldsymbol{h}_{v,t}^U = GRU\left(\boldsymbol{h}_{v,t-1}^U, \boldsymbol{x}_{v,t}^U\right), \tag{10}$$

其中$\boldsymbol{h}_{v,t}^U$代表视角v的隐藏状态，我们使用隐藏状态作为每个视图下学习到的偏好. 直接将隐藏状态送入*softmax*网络层就可以产生推荐结果，但由于*softmax*函数会使得网络更倾向于选择置信度最高的类别，而忽视其他可能的选择. 这种情况下，即使某些推荐结果并不准确，它们也会被模型过度自信地排在前面. 这可能导致用户接收到与其实际兴趣不符的推荐，降低推荐系统的可靠性和用户满意度. 因此，需要设计一种能够有效保障推荐结果可靠性的方法，以提升推荐性能.

3.3 可信决策

本文并没有像传统做法一样简单地将所有视角中的信息整合到一个表征中，而是在证据层面评估每种视角的置信度，最后整合不同视角的信息获得每个项目的推荐概率和当前预测的总体不确定性.

推荐本质上是一种具有$|I|$个类别的多分类问题. 只需把传统的分类器稍加改动转化为基于证据的分类器，神经网络可以从输入中捕捉证据来诱导分类意见[29]. 我们将传统分类器*softmax*层替换为激活函数层（即*ReLu*层）. 因此，对于第v个视图，这些非负值就被视为证据向量$\boldsymbol{e}_v = (e_{v,1}, e_{v,2}, \cdots, e_{v,|I|})$. 基于主观逻辑[30]的可信决策理论定义了一个基于证据获得不同类别概率（信念质量）和总体不确定性（不确定性质量）的理论框架. 证据与狄利克雷分布参数密切相关，即狄利克雷分布的参数$\alpha_{v,i}$是由证据$e_{v,i}$导出的：

$$\alpha_{v,i} = e_{v,i} + 1. \tag{11}$$

主观逻辑会为每个项目分配一个信念质量$b_{v,i}$，并为该视角下的整体框架分配一个总体不确定性c_v. 在视角v中，$|I|+1$个质量值是非负的且总和为1：

$$\sum_{i=1}^{|I|} b_{v,i} + c_v = 1, \tag{12}$$

其中$b_{v,i} \geqslant 0$表示第i个项目被推荐的概率，c_v代表该视角下的总体不确定性. 具体计算方式为

$$b_{v,i} = \frac{e_{v,i}}{D_v} = \frac{\alpha_{v,i} - 1}{D_v}, c_v = \frac{|I|}{D_v}, \tag{13}$$

其中$D_v = \sum_{i=1}^{|I|} \left(e_{v,i} + 1\right) = \sum_{i=1}^{|I|} \alpha_{v,i}$代表狄利克雷强度. 信念分配可以被视为主观意见，也就是说，项目i的证据越多，它被分配的概率也就越大.

不同视角的质量往往因样本质量不同而存在差异，对于不同视角间的融合，我们提出了一种自适应的融合方法，而不是给每个视角分配固定权重. Dempster-Shafer理论允许将不同来源的证据结合起来，在我们的算法模型中，它结合了多模态视角信息和商品ID视角信息的质量分配集，从而获得联合质量，可以灵活扩展到多个视图融合. 具体来说，$M_1 = \{\{b_{1,i}\}_{i=1}^{|I|}, c_1\}$代表多模态视角的质量集，$M_2 = \{\{b_{2,i}\}_{i=1}^{|I|}, c_2\}$代表商品ID视角的质量集，两者一起计算

可以得到联合质量集M:

$$M = M_1 \oplus M_2. \tag{14}$$

各部分更具体的计算规则为

$$b_i = \frac{1}{1-\beta}\left(b_{1,i}b_{2,i} + b_{1,i}c_2 + b_{2,i}c_1\right), c = \frac{c_1 c_2}{1-\beta}, \tag{15}$$

其中$\beta = \sum_{i \neq j} b_{1,i}b_{2,j}$代表2个质量集中的冲突程度, $\frac{1}{1-\beta}$用于归一化. 因此, 相应的联合证据和狄利克雷分布的参数可以诱导为

$$D = \frac{|I|}{c}, e_i = b_i \times D, \alpha_i = e_i + 1. \tag{16}$$

综上, 我们可以得到估计的联合证据$\boldsymbol{e}$和狄利克雷分布的相应参数$\boldsymbol{\alpha}$, 从而得出每个商品的最终概率和总体不确定性. 相应的损失函数由交叉熵损失调整得到:

$$\mathcal{L}_{\text{cer}}(\boldsymbol{\alpha}_v) = \sum_{j=1}^{|I|} y_{ij}\left(\psi(D_v) - \psi\left(\alpha_{v,j}\right)\right), \tag{17}$$

其中$\psi(\cdot)$是digamma函数. 损失函数允许每个样本的正确标签比其他类别标签产生更多的证据. 因此, 本文提出的算法模型总体损失函数为

$$L = \mathcal{L}_{\text{rec}}\left(S^U\right) + \lambda\left(\sum_{v=1}^{|V|}\mathcal{L}_{\text{cer}}(\boldsymbol{\alpha}_v) + \mathcal{L}_{\text{cer}}(\boldsymbol{\alpha})\right). \tag{18}$$

4 实 验

为了验证模型的有效性, 本文在4个真实公开的数据集上进行了实验. 首先介绍数据集、评估指标、基线方法和参数设置. 随后将模型的性能与基线的性能进行比较. 最后进行消融实验并回答以下问题:

问题1. 模型能否超越经典的序列推荐方法和多模态序列推荐方法.

问题2. 不同的组件如何影响模型的性能.

问题3. 超参数的设置对模型有何影响.

4.1 数据集

本文选择了4个真实世界的公共数据集进行实验, 包括Amazon-Beauty, Amazon-Sport, Amazon-Toys, Yelp.我们为所有数据集选择了产品图片和用户评论作为附加的多模态信息. 参照以前的方法[31]对数据进行预处理, 保留5次及5次以上交互的用户和项目. 然后, 根据时间戳对每个用户的交互进行排序, 所有交互都被视为隐式反馈. 4组数据的统计结果如表1所示.

Table 1 Statistical Results of Four Datasets

表1 4个数据集的统计结果

数据集	用户	项目	交互	稀疏度/%
Amazon-Beauty	22 363	12 101	198 502	99.93
Amazon-Sport	35 598	18 357	2 563 083	99.91
Amazon-Toys	19 412	11 924	167 597	99.93
Yelp	30 499	20 068	2 563 083	99.95

4.2 评估指标

为了评估序列推荐系统的性能, 本文采用Top-k召回率(*Recall*@k)和Top-k归一化折扣累积增益(*NDCG*@k)两种评价指标, 值越高, 表示模型的性能越好, 其中$k \in \{10, 20\}$. *Recall*的计算公式为

$$Recall = \frac{|R(u) \bigcap T(u)|}{T(u)}, \tag{19}$$

其中$R(u)$表示模型预测出需要推荐给用户的商品集合, $T(u)$表示真实测试集中被推荐的集合. *NDCG*的计算公式为

$$NDCG = \frac{DCG}{IDCG}. \tag{20}$$

式(20)中各部分的详细计算公式为

$$DCG = \sum_{i=1}^{p}\frac{2^{rel_i}-1}{\text{lb}(i+1)}, IDCG = \sum_{i=1}^{|REL|}\frac{2^{rel_i}-1}{\text{lb}(i+1)}. \tag{21}$$

其中rel_i表示第i个位置商品的相关性得分, $|REL|$表示理想排序下第i个位置的相关性得分. 根据文献[32]的建议, 我们以完整排名的方式评估模型性能以进行公平比较. 排名结果来自整个项目集, 而不是从样本集上获得.

4.3 基线方法

我们选择了2类方法: 经典的序列推荐方法(即GRU4Rec, SASRec, Caser, BERT4Rec); 结合不同侧信息的新方法, 其中包括利用商品属性的DIF-SR、利用评论信息的RNS和利用多模态信息的MMSRec. 以上提到的方法都只考虑了与项目相关的侧信息.

1)GRU4Rec[3]. 基于会话的推荐模型, 使用门控循环单元更好地捕获用户序列行为, 提高推荐性能.

2)SASRec[31]. 序列推荐模型, 利用自我注意力机制分析整个用户行为序列, 预测下一个可能交互的商品.

3)Caser[33]. 序列推荐模型, 采用卷积滤波器捕获全局级别和局部级别的用户序列行为模式.

4)BERT4Rec[34]. 具有双向编码器的序列推荐模型, 它使用类似完形填空任务去训练双向编码器并预测被屏蔽的项目.

5)DIF-SR [16]. 考虑商品各种属性的序列推荐模型, 它将和商品相关的侧信息从输入层转移至注意力层, 并将注意力计算与项目表征分离开来.

6）RNS[35]. 文本评论驱动的序列推荐模型，综合考虑了用户的内在偏好和序列行为模式.

7）MMSRec[36]. 自监督的多模态模型，整合了视觉模态和文本模态的特征，采用双塔架构和自监督学习来提高序列推荐性能.

4.4 参数设置

对于 4.3 节提到的模型，我们都使用了公开的源代码和真实世界的数据集. 模型中用到的超参数，我们遵循了原作者的建议，列出了每个基线方法在建议超参数设置下的最优结果. 本文提出的模型 Large-TR 基于 pytorch 实现，大语言模型为 qwen-max-1201，使用自适应梯度算法进行了 100 次训练，批样本大小为 32，学习率为 0.001. 对于其他超参数，我们遍历所有参数设置以找便到最优结果，包括：隐藏层大小的范围是{100, 200, 400, 500, 600, 700, 1 000}，层数的范围是{1, 2, 3, 4}，图像嵌入维数的范围是{256, 512, 1 024}，文本嵌入维数的范围是{256, 512, 1 024}. 不同数据集的结果可能会有所差异.

4.5 总体性能（问题 1）

在 4 个数据集上的实验结果如表 2 所示. 通过这些结果，我们可以观察到模型 Large-TR 在多模态序列推荐中展示出了优异的性能. 在大多数情况下，包含辅助信息的推荐方法（如 MMSRec, RNS, DIF-SR）在性能表现上优于传统的推荐方法（如 GRU4Rec, Caser, BERT4Rec, SASRec），这验证了引入辅助信息提升性能的可行性. 此外，MMSRec 相较于其他引入辅助信息的方法（如 RNS, DIF-SR）表现更为出色，这表明使用多模态内容的方法与仅使用单一模态的推荐方法相比，能让模型从更多维度和更丰富的上下文中提取信息，从而更准确地反映用户的真实意图和兴趣偏好. 最后，Large-TR 几乎在所有基线方法中均取得了最优表现. 这主要得益于 3 个方面：首先，我们充分利用了用户生成的多模态内容，这使得我们的算法模型具有更丰富和多样化的输入特征. 其次，我们特别考虑了多模态内容中存在的噪声问题，设计了基于大语言模型的降噪策略，提高了数据质量. 最后，通过可信决策机制，我们从多模态信息和项目信息 2 个视角获取一致且可信的推荐决策.

4.6 消融实验（问题 2）

为了确定各种类型的辅助信息对模型 Large-TR 的影响，我们使用不同的辅助信息进行了消融实验. 我们采用在推荐平台上常见的数据：项目 ID、用户评论（文本信息）、商品图片（视觉信息）. 针对以上数据，我们分别设计了 4 种类型的实验如表 3 所示.

表 3 表明，仅使用文本或图片信息就能提高推荐性能，这证明了侧面信息的有效性. 此外，2 种模态信

Table 2 Experimental Results for Each Dataset

表 2 各数据集的实验结果 %

k	模型	Amazon-Beauty		Amazon-Sport		Amazon-Toys		Yelp	
		Recall	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*
10	GRU4Rec	5.29	2.66	3.12	1.57	3.70	1.84	3.61	1.84
	Caser	4.74	2.39	2.27	1.18	3.61	1.86	3.80	1.97
	BERT4Rec	5.29	2.37	2.95	1.30	5.33	2.34	5.24	3.27
	SASRec	8.28	3.71	5.26	2.33	8.31	3.75	6.50	4.01
	RNS	8.96	4.04	5.32	2.35	9.36	4.25	6.76	4.12
	DIF-SR	9.08	4.46	5.56	2.64	10.13	5.04	6.98	4.19
	MMSRec	**9.49**	4.76	6.35	3.23	11.54	6.14	7.49	4.52
	Large-TR（本文）	9.43	**4.85**	**6.39**	**3.35**	**11.83**	**6.92**	**8.43**	**4.90**
20	GRU4Rec	8.93	3.44	4.82	2.01	5.88	2.39	5.92	2.43
	Caser	7.31	3.02	3.64	1.53	5.66	02.38	6.08	2.55
	BERT4Rec	8.15	3.09	4.65	1.73	7.87	2.97	7.56	3.85
	SASRec	11.97	4.64	7.73	2.95	11.68	4.60	9.28	4.71
	RNS	12.31	4.78	7.74	2.99	12.31	5.12	9.54	4.72
	DIF-SR	12.84	5.41	7.98	3.25	13.82	4.97	10.03	2.96
	MMSRec	13.41	5.77	9.57	4.23	10.03	7.04	11.07	5.25
	Large-TR（本文）	**14.34**	**6.22**	**10.18**	**4.72**	**11.68**	**7.83**	**12.69**	**5.95**

注：加粗的数字代表最佳实验结果.

Table 3 Side-Information Ablation Results for Each Dataset

表 3 各数据集的辅助信息消融结果 %

辅助信息	Amazon-Beauty		Amazon-Sport	
	Recall@10	*NDCG*@10	*Recall*@10	*NDCG*@10
项目 ID	5.29	2.66	3.12	1.57
项目 ID+图片	7.21	3.78	4.84	2.39
项目 ID+文本	8.93	4.49	5.50	2.67
全部	9.43	4.85	6.39	3.35

息的结合也能进一步提高推荐性能, 这表明我们提出的融合算法可以有效地利用来自不同模态的信息.

此外, 为了研究模型 Large-TR 中不同组件对实验结果的影响, 我们设计了相应的消融研究, 并分别在 Amazon-Beauty, Amazon-Toys, Yelp 数据集上进行了实验, 实验结果如表 4 所示.

相关实验设置如下:

T1. Large-TR w/o LLM: 图片和文本经过对应的编码器后直接拼接, 其他部分不变;

T2. Large-TR w/o Trust: 可信决策部分由单层 *softmax*代替, 其他部分不变;

T3. Large-TR w/o LLM&Trust: 包含所有组件的模型.

由表 4 可以得出以下结论: 引入大语言模型去噪和可信决策 2 个组件可以有效提升推荐算法的性能, 这说明我们模型设计是合理且有效的. 大语言模型利用其丰富的语义知识, 过滤不相关的图文内容, 显著提高了多模态数据质量. 最后, 可信决策机制在证据层面评估每个视角的置信度, 整合不同视角的信息获得每个项目的推荐概率和当前预测的总体不确定性, 提供可信推荐.

4.7 参数设置 (问题 3)

本节主要研究了网络的隐藏层维度和层数这 2 个超参数对实验结果的影响, 图 3 显示了模型 Large-TR 在不同隐藏层维度设置下的 *Recall*@10 和 *NDCG*@10 分数. 在 2 个数据集的结果表明, 当隐藏层维度为 600 时, 模型性能最佳. 图 4 显示了不同网络层数与模型性能的关系, 可以看出网络层数为 2 时, 模型效果最好.

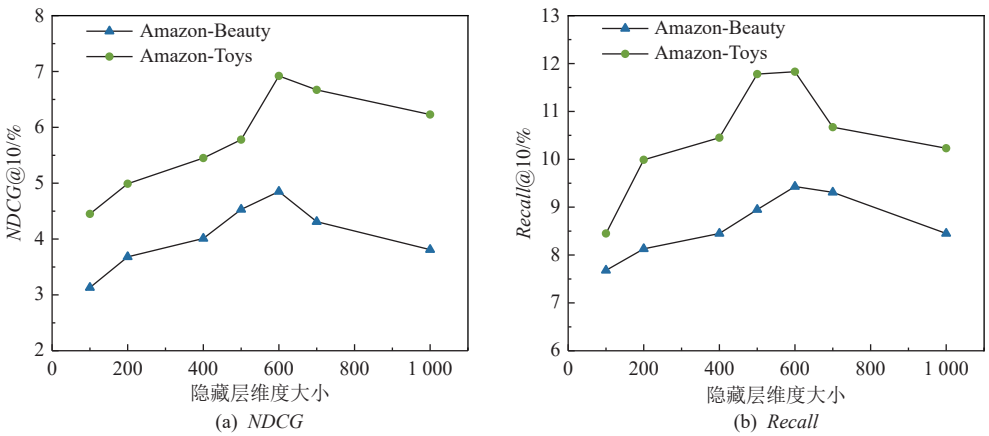
5 总 结

本文探讨了基于多模态内容的序列推荐问题, 针对如何从噪声较多的用户生成内容中获得可信推荐这一挑战, 提出了 Large-TR 算法. 该算法利用大语言模型的丰富语义知识, 识别和过滤用户生成内容中的噪声, 提高多模态数据质量. 接着, 从商品 ID 和多模态信息 2 个视角挖掘用户偏好, 动态评估推荐

Table 4 Component Ablation Results for Each Dataset

表 4 各数据集的组件消融结果 %

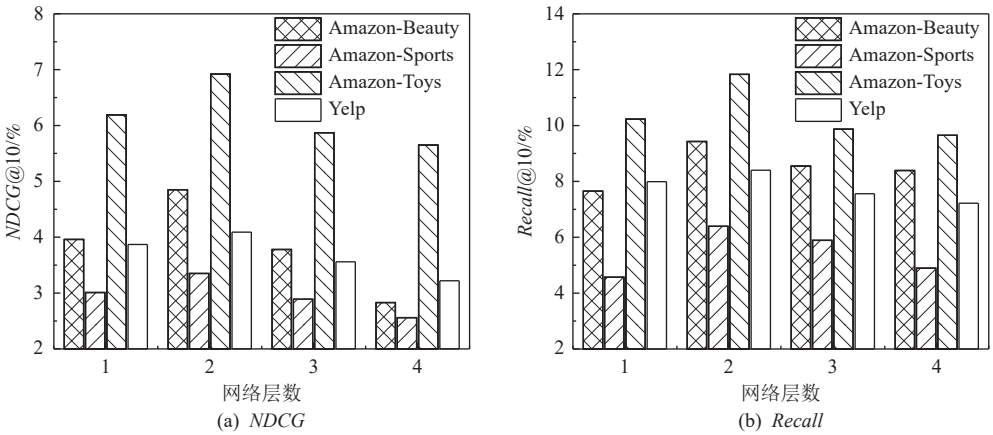
实验	Amazon-Beauty		Amazon-Toys		Yelp	
	Recall@10	*NDCG*@10	*Recall*@10	*NDCG*@10	*Recall*@10	*NDCG*@10
T1	8.82	4.49	9.36	4.25	7.11	4.32
T2	9.11	4.67	10.88	5.98	8.01	4.73
T3	9.43	4.85	11.83	6.92	8.43	4.90



(a) *NDCG* (b) *Recall*

Fig. 3 Impact of hidden layer dimension on performance

图 3 隐藏层维度对性能的影响



(a) *NDCG* (b) *Recall*

Fig. 4 Impact of network layers on performance

图 4 网络层数对性能的影响

结果的不确定性，并提供相应的推荐结果及置信度. 在 4 个公开数据集上的实验结果表明，该模型在性能上优于大多数现有的序列推荐模型. 最后的消融实验证明了 Large-TR 中各个组件的有效性. 未来的工作会考虑可解释的可信推荐系统，帮助用户更好地理解推荐结果. 此外，对置信度的评估也是一个值得深入研究的方向.

作者贡献声明：闫萌提出了算法思路和实验方案，并撰写论文；徐偲辅助写作并修改论文；黄海槟负责数据处理、实验实施及图表绘制；赵伟与管子玉提供了关键的指导意见.

参 考 文 献

[1] Wu Bin, He Xiangnan, Sun Zhongchuan, et al. ATM: An attentive translation model for next-item recommendation[J]. IEEE Transactions on Industrial Informatics, 2019, 16(3): 1448−1459

[2] McAuley J, Targett C, Shi Qinfeng, et al. Image-based recommendations on styles and substitutes [C] // Proc of the 38th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2015: 43−52

[3] Hidasi B, Karatzoglou A, Baltrunas L, et al. Session-based recommendations with recurrent neural networks[J]. arXiv preprint, arXiv: 1511.06939, 2015

[4] Chang Jianxin, Gao Chen, Zheng Yu, et al. Sequential recommendation with graph neural networks [C] // Proc of the 44th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2021: 378−387

[5] Chen Xu, Chen Hanxiong, Xu Hongteng, et al. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation [C] // Proc of the 42nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2019: 765−774

[6] Cheng Zhiyong, Chang Xiaojun, Zhu Lei, et al. MMALFM: Explainable recommendation by leveraging reviews and images[J]. ACM Transactions on Information Systems, 2019, 37(2): 1−28

[7] Chu Weita, Tsai Y. A hybrid recommendation system considering visual information for predicting favorite restaurants[J]. World Wide Web, 2017, 20: 1313−1331

[8] Qian Shengsheng, Zhang Tianzhu, Xu Changsheng. Multi-modal multi-view topic-opinion mining for social event analysis [C] // Proc of the 24th ACM Int Conf on Multimedia. New York: ACM, 2016: 2−11

[9] Chen Changrui, Han Jungong, Debattista K. Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(8): 5595−5611

[10] Silva A. Enhancing deep multimodal representation: Online, noise-robust and unsupervised learning [D]. Parkville, AU: University of Melbourne, 2022

[11] He Ruining, McAuley J. Fusing similarity models with Markov chains for sparse sequential recommendation [C] // Proc of the 16th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2016: 191−200

[12] Kabbur S, Ning Xia, Karypis G. FISM: Factored item similarity models for top-*N* recommender systems [C] // Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 659−667

[13] Rendle S. Factorization machines [C] // Proc of the 10th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2010: 995−1000

[14] Zimdars A, Chickering M, Meek C. Using temporal data for making recommendations [C] // Proc of the 17th Conf in Uncertainty in Artificial Intelligence. New York: ACM, 2001: 580–588

[15] Zhou Kun, Wang Hui, Zhao Wayne Xin, et al. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization [C] // Proc of the 29th ACM Int Conf on Information & Knowledge Management. New York: ACM, 2020: 1893−1902

[16] Xie Yueqi, Zhou Peilin, Kim S. Decoupled side information fusion for sequential recommendation [C] // Proc of the 45th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2022: 1611−1621

[17] He Ruining, McAuley J. VBPR: Visual Bayesian personalized ranking from implicit feedback [C] // Proc of the 29th AAAI Conf on Artificial Intelligence. Palo Alto, CA, 2015: 144−150

[18] Lei Chenyi, Liu Yong, Zhang Lingzi, et al. Semi: A sequential multi-modal information transfer network for e-commerce micro-video recommendations [C] // Proc of the 27th ACM SIGKDD Conf on Knowledge Discovery & Data Mining. New York: ACM, 2021: 3161−3171

[19] Geng Shijie, Liu Shuchang, Fu Zuohui, et al. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (p5) [C] // Proc of the 16th ACM Conf on Recommender Systems. New York: ACM, 2022: 299−315

[20] Lin Junyang, Men Rui, Yang An, et al. M6: Multi-modality-to-multi-modality multitask mega-transformer for unified pretraining [C] // Proc of the 27th ACM SIGKDD Conf on Knowledge Discovery & Data Mining. New York: ACM, 2021: 3251−3261

[21] Dai Sunhao, Shao Ninglu, Zhao Haiyuan, et al. Uncovering ChatGPT's capabilities in recommender systems [C] // Proc of the 17th ACM Conf on Recommender Systems. New York: ACM, 2023: 1126−1132

[22] Bao Keqin, Zhang Jizhi, Zhang Yang, et al. Tallrec: An effective and efficient tuning framework to align large language model with recommendation [C] // Proc of the 17th ACM Conf on Recommender Systems. New York: ACM, 2023: 1007−1014

[23] Lin Xinyu, Wang Wenjie, Li Yongqi, et al. A multi-facet paradigm to bridge large language model and recommendation[J]. arXiv preprint, arXiv: 2310.06491, 2023

[24] Mann B, Ryder N, Subbiah M, et al. Language models are few-shot learners[J]. arXiv preprint, arXiv: 2005.14165, 2020

[25] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, arXiv: 1810.04805, 2018

[26] Zhang Qi, Wang Jiawen, Huang Haoran, et al. Hashtag recommendation for multimodal microblog using co-attention network [C] // Proc of the 26th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2017: 3420−3426

[27] Radford A, Kim W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C] // Proc of the 24th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[28] Zhang Ying, Lu Huchuan. Deep cross-modal projection learning for image-text matching [C] // Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 686−701

[29] Xu Cai, Si Jiajun, Guan Ziyu, et al. Reliable conflictive multi-view learning[C] //Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 16129−16137

[30] Jsang A. Subjective Logic: A Formalism for Reasoning Under Uncertainty[M]. Berlin: Springer, 2018

[31] Kang Wangcheng, McAuley J. Self-attentive sequential recommendation [C] // Proc of the 18th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2018: 197−206

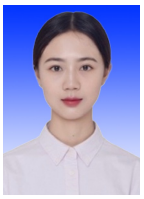
[32] Krichene W, Rendle S. On sampled metrics for item recommendation [C] // Proc of the 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2020: 1748−1757

[33] Tang Jiaxi, Wang Ke. Personalized top-N sequential recommendation via convolutional sequence embedding [C] // Proc of the 11th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2018: 565−573

[34] Sun Fei, Liu Jun, Wu Jian, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer [C] // Proc of the 28th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2019: 1441−1450

[35] Li Chenliang, Niu Xichuan, Luo Xiangyang, et al. A review-driven neural model for sequential recommendation[J]. arXiv preprint, arXiv: 1907.00590, 2019

[36] Song Kunzhe, Sun Qingfeng, Xu Can, et al. Self-supervised multi-modal sequential recommendation[J]. arXiv preprint, arXiv: 2304.13277, 2023



Yan Meng, born in 1997. PhD candidate. Her main research interests include multi-modal learning, recommender system, and knowledge graph.

闫　萌, 1997 年生. 博士研究生. 主要研究方向为多模态学习、推荐系统、知识图谱.



Xu Cai, born in 1994. PhD, associate professor, master supervisor. His main research interests include trustworthy learning and multi-view learning.

徐　偲, 1994 年生. 博士, 副教授, 硕士生导师. 主要研究方向为可信学习、多视角学习.



Huang Haibin, born in 2000. Master candidate. His main research interests include recommender system and multi-modal learning.

黄海槟, 2000 年生. 硕士研究生. 主要研究方向为推荐系统、多模态学习.



Zhao Wei, born in 1979. PhD, professor, PhD supervisor. His main research interests include signal processing, pattern recognition, and intelligent systems.

赵　伟, 1979 年生. 博士, 教授, 博士生导师. 主要研究方向为信号处理、模式识别、智能系统.



Guan Ziyu, born in 1982. PhD, professor, PhD supervisor. His main research interests include attributed graph mining, expertise modeling, and recommender system.

管子玉, 1982 年生. 博士, 教授, 博士生导师. 主要研究方向为属性图挖掘、专家知识建模、推荐系统.

附录 A.

在本附录里，我们提供了提示词的优化过程. 本文从初始的提示符和任务描述开始，提供少量的样本示例. 然后，通过这些样本校准优化提示符，生成新的提示符. 在这个过程中，系统会执行以下步骤：

1) 为任务和当前提示符提出一些具有挑战性和多样性的样本；

2) 在生成的样本上评估当前提示符，并进行分析；

3) 根据评估结果，修改提示符提高分数. 当分数没有提高后，优化过程就会结束.

具体而言，我们对本文的提示符进行的优化迭代过程如下所示：

A0. 描述问题，要求给出结果.

Input Template: I have an image {item_image} and a text {user_review}, please tell me if they are similar.

A1. 指定输入样本的主体.

Input Template: I have an image of item {item_image} and a review text of item {user_review}, please tell me if they are similar.

A2. 新增输出引导的模版.

Input Template: I have an image of item {item_image} and a review text of item {user_review}, please tell me if they are similar.

Output Template: Yes or No.

A3: 重新描述任务，并给定输出模版.

Input Template: I have an image of item {item_image} and a review text of item {user_review}, the task is judge if they similar, please tell me if they are similar.

Output Template: Yes or No.

A4: 更改输出引导的模板.

Input Template: It is an image of the item {item_image}, and it is a text review{user_review}. The task is judge if they match, answer match or non-match.

Output Template: match or non-match.

A5: 新增场景化任务描述.

Input Template: Given a user text comment and an image of a product on an e-commerce website, it is an image of the item {item_image}, and it is a text review {user_review}. Determine if the descriptions of the text and the image match, answer match or non-match.

Output Template: match or non-match.