计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(2):63−74 [doi: 10.15888/j.cnki.csa.008940]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于全连接张量网络的多模态与多样性推荐算法①

孟诗蓓[1], 郑　睿[2], 常　亮[1], 陈玉珑[3], 孟睿伟[4], 程　诺[4]

[1](北京师范大学 人工智能学院, 北京 100875)
[2](北京师范大学 数学科学学院, 北京 100875)
[3](上海交通大学, 上海 200240)
[4](中国国家博物馆, 北京 100006)
通信作者: 常　亮, E-mail: changliang@bnu.edu.cn

摘　要: 在全媒体时代下, 基于多模态数据的推荐具有重要意义. 本文使用文本、音频、图像 3 种模态数据进行推荐, 通过两个阶段进行张量融合: 第 1 阶段通过 3 个平行分支对任意两个模式的相关性进行建模和融合, 第 2 阶段再将 3 个分支的结果进行融合, 不仅考虑了两模态之间的局部交互作用, 并且消除了模态融合顺序对结果的影响; 在推荐模块中, 将融合特征通过堆叠降噪自编码器作为协同过滤的辅助特征进行推荐. 本文所构建的推荐系统中模态融合与推荐采用端到端的训练过程. 同时, 为了解决推荐结果中存在的相似度高、多样性差的问题, 我们基于二阶段的张量模态融合特征构建相似度矩阵, 在已有推荐结果的基础上进一步精化结果, 实现快速的多样性推荐. 实验证明, 基于本文提出的多模态融合特征的推荐模型不仅能够有效地提升推荐性能, 并且能够增强推荐结果的多样性.

关键词: 张量网络; 多模态融合; 多样性推荐; 堆叠降噪自编码器; 协同过滤; 推荐算法

引用格式: 孟诗蓓,郑睿,常亮,陈玉珑,孟睿伟,程诺.基于全连接张量网络的多模态与多样性推荐算法.计算机系统应用,2023,32(2):63–74. http://www.c-s-a.org.cn/1003-3254/8940.html

Multimodal and Diverse Recommendation Algorithm Based on Fully-connected Tensor Networks

MENG Shi-Bei[1], ZHENG Rui[2], CHANG Liang[1], CHEN Yu-Long[3], MENG Rui-Wei[4], CHENG Nuo[4]

[1](School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China)
[2](School of Mathematical Sciences, Beijing Normal University, Beijing 100875, China)
[3](Shanghai Jiao Tong University, Shanghai 200240, China)
[4](National Museum of China, Beijing 100006, China)

Abstract: In the all-media era, recommendation based on multimodal data is of great significance. This study proposes recommendation based on data in three modalities: text, audio, and image. Tensor fusion is implemented in two stages: The correlation between any two modes is modeled and fused by three parallel branches in the former stage, and the results of the three branches are then fused in the latter stage. This approach not only considers the local interaction between two modalities but also eliminates the influence of the modality fusion order on the result. In the recommendation module, the fused features are input to the stacked denoising auto-encoder and are then used as auxiliary features of collaborative filtering for recommendation. In the recommendation system constructed, an end-to-end training process is adopted for modality fusion and recommendation. Moreover, to overcome the high similarity and poor diversity of the recommendation results, this study also constructs a similarity matrix with the fused features of the tensor modalities in the two stages to further refine the results on the basis of the available recommendation results and thereby achieve rapid

① 基金项目: 国家重点研发计划 (2019YFC1521100); 国家自然科学基金 (61977063); 国家自然科学基金重大项目 (72192821); 上海市科委重大项目 (21511101200)
收稿时间: 2022-06-20; 修改时间: 2022-07-08; 采用时间: 2022-08-15; csa 在线出版时间: 2022-10-28
CNKI 网络首发时间: 2022-11-16

diversified recommendation. The experimental results show that the recommendation model based on the proposed multimodal fused features can not only effectively improve recommendation performance but also enhance the diversity of recommendation results.

Key words: tensor network; multimodal fusion; diversity recommendation; stacked denoising auto-encoder; collaborative filtering; recommendation algorithm

当今, 随着全媒体时代的发展, 一些音视频网站与应用正在逐渐盛行, 例如腾讯视频、爱奇艺、TikTok、快手等, 人们正生活在一个多种媒体相互作用、互为补充的社会里. 这些应用的特点是: 一方面, 它们提供了大量的物品数据, 其中包括文本、图像、音频、视频等不同媒体类型的物品信息; 另一方面, 这些应用的用户粘性很大程度上取决于推荐系统的好坏. 因此, 能够融合多类媒体和多种信息进行更加精准、更为个性化的推荐具有非常重要的意义.

传统的推荐算法多使用用户历史行为信息 (例如: 点赞、评分、点击量等) 对用户和物品的交互通过偏好程度的相似度进行建模, 这种方法也被称作协同过滤算法. 但随着实际的应用, 这个算法一方面被认为不能够很好地解决稀疏性和冷启动的问题[1]; 另一方面, 如果加入辅助信息, 能够丰富物品的表达, 在引入神经网络建模时, 将能够建立更多用户和物品特征的交互, 使推荐系统的性能大大提高[2].

当前国内外学者对于推荐算法的研究中, 根据辅助信息的种类, 其工作可以被分为 3 大类: 基于文本的推荐算法、基于视觉模态的推荐算法和基于多模态的推荐算法.

文献 [3,4] 是基于文本信息的推荐算法, Wei 等[3] 提出基于紧耦合深度协同过滤的推荐模型, 使用改进的分解模型 Time-SVD++, 并使用多个自编码器学习物品的文本特征, 能够有效解决冷启动的问题. Frolov 等[4] 提出一种集成的混合算法, 扩展传统的奇异值分解方法, 将交互数据和辅助的文本信息进行联合分解, 同样能够很好地解决冷启动的问题.

文献 [5–7] 是基于图像信息的多模态推荐算法, Lei 等[5] 提出双网深度网络用于图像的推荐, 将图片和用户偏好使用两个子网络映射到同一个潜在语义空间中, 形成了更有效的偏好表示和图像表示; Tang 等[6] 对视觉推荐模型的健壮性进行研究, 提出 AMR 模型, 通过对抗学习生成鲁棒性更强的推荐模型, 并验证了其对于图像推荐和视觉感知产品推荐的有效性; Qiu 等[7] 提出 CausalRec 模型, 通过因果推理框架仅保留有效的视觉特征, 解决了现有推荐系统存在的视觉偏差问题.

文献 [8–10] 是基于多模态融合的推荐算法, Oramas 等[8] 提出使用深度网络架构将文本和音频信息与用户反馈数据相结合, 使用简单的多层感知机对级联特征进行融合. Sun 等 [9] 提出了一种用于融合文本与图像异构模态的紧耦合深度网络模型, 网络将原始图片与文本信息作为输入, 从特征提取级别开始训练, 在训练过程中同时对特征提取模块和协同过滤模型进行优化, 得到了较好的效果. 肖庆华等 [10] 则针对互补推荐的目标, 提出基于图片、文本以及评分的多模态互补物品特征提取算法: 结合卷积神经网络、文本向量化、贝叶斯推断 3 种方法, 提高了推荐系统的准确率, 同时使用 Bandits 算法提高了推荐系统的多样性.

与此同时, 文本、图像与音频 3 种模态特征生成融合表征已被国内外学者广泛关注: Huang 等[11] 提出视觉-语言的 Transformer 模型, 将上述 3 种模态特征应用于跨媒体检索; Wang 等[12] 受到人类记忆的再建构和联想性质整合启发, 将 3 种模态作为自编码器的输入来解决学习多模态单词表示的问题; 而 Zadeh 等[13] 则从情感分析的角度比较了多模态协同学习和单峰学习的区别. 在实际的系统应用方面, 中科院自动化研究所发布了依托武汉人工智能计算中心算力研发的跨模态通用人工智能平台“紫东太初”, 在该平台上, Liu 等[14] 首次提出了视觉-文本-语音三模态预训练模型, 实现了三模态间的相互转换和生成.

目前, 将文本、图像与音频的融合与推荐系统结合应用目前尚未得到广泛应用. 而音频数据对于当今时代的推荐系统有着重要的作用, 声音中所包含的环境声、人物互动的情绪语气、甚至是其中包含的背景音乐等都是非常重要的物品风格, 而这些特征是文本和图像无法表征的[15]; 并且, 对于拥有多个模态的推荐系统来说, 线性模型不足以表示复杂的相互关系, 一次性融合所有

特征将忽略复杂的局部相关性, 而只考虑双线性池化时, 需要先对两个模态融合, 再将融合结果与第 3 个模态进行融合, 其模型表达能力可能也会受到交互顺序的影响. 因此, 在基于多模态的方法中, 探索多模态数据的异质性, 提高模型的泛化能力, 仍然是一个重大的挑战.

本文基于上述问题, 提出基于张量网络的多模态推荐算法. 图像中的显著性区域和文本中的关键性单词具有较强的语义相关性[16], 另一方面, 张量融合方法通过张量外积将输入的多个模态转化为一个高维张量, 再将其映射回一个低维输出向量空间, 通过这种方式能够计算不同模态元素之间的相关性, 从而对跨模态之间的交互关系进行建模[17], 与早期融合在输入级别上简单地连接多模态特征相比, 能够更有效地建立模态内部之间的交互关系[18], 同时又比注意力机制等复杂的网络模型更为简单有效. 因此, 在高维张量上进行模态融合被认为是一种有效的模态交互方式. 本文提出的模型将首先在浅层的显著特征上通过 3 个分支对交互性进行两两融合建模, 对于该过程中维度过高的张量使用低秩矩阵因子对其进行降维; 之后将 3 个分支中的特征再次进行张量融合, 得到第 2 阶段的融合特征; 将二阶段张量融合后的特征输入深度协同过滤模型中, 通过联合损失函数对结果进行训练. 整个网络是一个端到端的训练过程, 能够大大提高的模型的表达能力. 近年来, 量子计算引起了研究者的广泛重视, 量子人工智能有望成为一种通向第 3 代人工智能的途径. 通过张量法可将参数矩阵整理为量子多体, 使用张量分解压缩模型参数, 可以大幅度提升计算资源的利用率, 降低计算的复杂度. 因此, 基于张量分解的推荐算法的量子计算模拟具有很好的发展前景[19].

为了最大化地利用融合特征, 本文使用融合特征在推荐结果的多样性上进行精化. 如今, 在越来越注重推荐准确率的同时, 推荐列表的多样性成为衡量推荐系统好坏的重要依据. 因此, 本文将模态融合特征用于多样性推荐中, 能够在不影响推荐准确率的前提下提高推荐系统的多样性, 使融合特征得到充分利用的同时, 能够完善整个推荐系统.

综上, 本文的贡献如下.

(1) 提出一个基于全连接张量网络的端到端训练的多模态推荐模型, 通过张量积的方式将级联特征映射为高维张量, 对不同模态的相关性通过张量融合进行建模, 提取最有效的融合特征用于深度协同过滤模型.

(2) 所建立的多模态融合的过程分成两个阶段, 第 1 阶段通过 3 个平行的分支对浅层多模态特征进行浅层特征融合, 第 2 阶段在深层特征上对 3 个分支进行融合. 充分考虑了不同模态的相关性, 以得到更好的融合特征表示.

(3) 在模态融合之后加入多样性推荐模块, 能够最大化地利用融合特征并构建出更完整的推荐算法.

(4) 在 MovieLens 数据集上构建文本、图像和音频的多模态数据集并进行实验, 取得了比基准实验更好的结果.

1 本文模型

1.1 模型框架

本文使用二阶张量池化的方法学习文本、图像、音频 3 种模态的交互关系并对其进行特征融合, 同时结合用户偏好矩阵, 与深度协同过滤分解方法结合, 设计端到端的网络用于多模态推荐, 模型框架如图 1 所示. 本方法的创新性在于: (1) 传统深度协同过滤模型仅考虑单一文本模态, 本文提出了针对图像、文本、音频 3 种模态的推荐算法; (2) 在多模态融合中, 使用了二阶段张量融合方法, 不仅注重模态的两两交互, 并且能够使特征不受到融合顺序的影响; (3) 模态融合之后使用联合优化方法, 构建端到端训练模型, 能够使得融合模态更加适用于推荐应用; (4) 对前一阶段的推荐结果进行基于行列式点过程的精化, 生成多样化推荐.

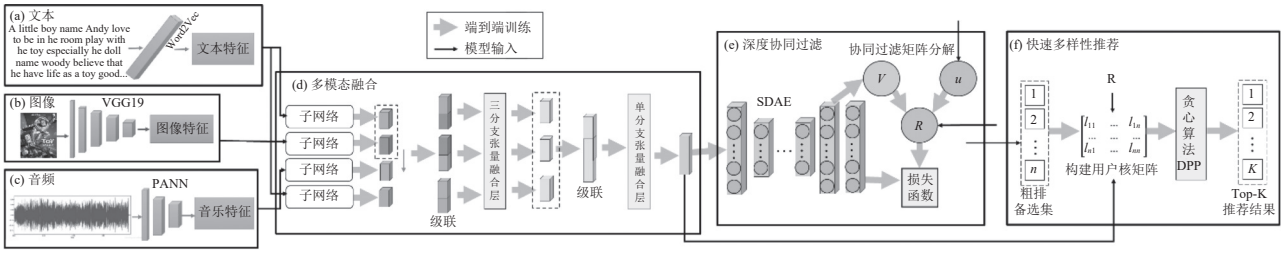

图 1 本文模型框架图

预处理阶段首先需要对文本、图像、音频 3 种模态特征进行提取. 为了使后续阶段能够更好地探索模态之间的交互性, 这里需要提取到更原始的具有充分异质性的模态特征, 因此本文分别使用针对不同模态的特征提取方法进行特征提取: 文本模态参考文献 [20] 中的语义表示模型提取 300 维文本特征, 视觉模态使用大规模预训练模型 VGG19[21] 提取 1 000 维视觉特征, 音频模态使用音频模式识别领域的大规模预训练模型 PANN[22] 提取 2 048 维声音模态特征.

本文模型主要包括 3 个部分: 基于张量网络的特征融合 (图 1(d)), 深度协同过滤推荐 (图 1(e)) 基于融合特征的多样性推荐 (图 1(f)). 模态融合模块我们参考文献 [15] 中的多模态融合方法, 为确保融合模块的特征不会受到交互顺序的影响, 并充分探索模态特征两两之间的交互模式, 使用 3 个平行的分支和两个先后的阶段: 首先将 3 个模态特征进行两两组合形成 3 个平行分支, 然后通过多阶张量积的方式将模态相乘, 以求得到更好的交互效果, 之后通过张量网络降维方法, 与低秩矩阵因子相乘将参数维度变低, 最后通过一个神经网络层将该阶段的结果输出; 第 2 个阶段再对 3 个分支的交互性进行建模: 与第 1 阶段的流程大体相同, 首先将 3 个分支的输出结果进行级联, 再对特征进行多阶张量相乘, 然后通过低秩矩阵分解将维度变低, 最后通过一个全连接层将最终融合特征输出. 对于深度协同过滤推荐模块: 我们将融合特征作为协同过滤推荐的辅助特征输入堆叠自编码器中, 在自编码器的中间层输出隐语义特征, 辅助协同过滤分解, 并在损失函数上做调整, 使得堆叠自编码器模型能够与协同过滤矩阵分解两个模块互相适应, 以更好地提高模型的鲁棒性.

1.2 模态融合模块

该模块的任务是通过建模模态之间的交互作用, 将单模态稀疏表示的高阶矩阵转变为紧凑表示的低阶矩阵[15]. 如图 2 所示: 分为 3 个阶段, 首先通过子网络自适应降维, 使用神经网络增强模型的非线性程度, 神经网络结构如表 1 所示, 将生成的高维特征维数降低; 再通过 3 个平行的分支对于 3 个模态两两之间的交互关系进行建模, 最后再对 3 个分支进行融合.

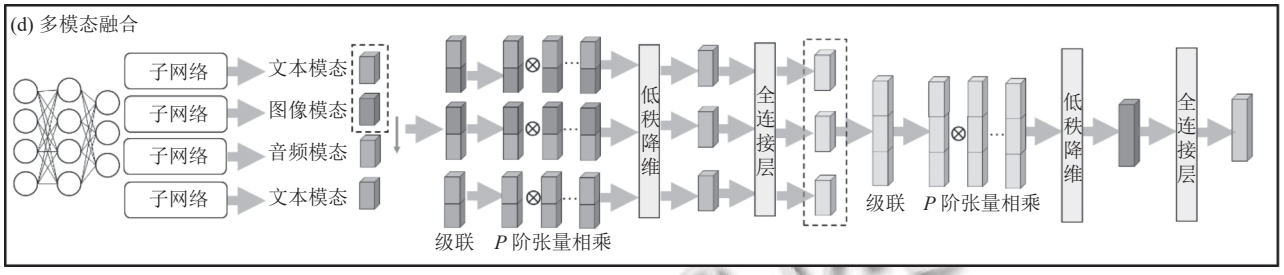

图 2 模态融合模块流程图

表 1 特征映射子网络结构表

层	激活函数
Batch_Normalization	—
Layer 1	ReLU
Layer 2	ReLU
Dropout	—
Layer 3	ReLU

对于第 1 个阶段: 首先将 3 个模态进行两两组合, 共有 3 种组合方式, 分别为音频和文本、音频和图像、文本和图像, 可以形成 3 个平行分支. 再对单个分支上的两个特征进行级联, 形成级联特征向量:

$$f_i = Concat(M_a, M_b) \tag{1}$$

其中, f_i表示第i个分支的级联特征, M 的不同角标a、b分别表示该分支上两个不同的模态特征. 之后, 为了使得级联特征之间拥有更好的交互性, 使用张量积的形式对级联特征向量进行P阶外积, 通过乘积形成高维张量, 以使模态之间进行充分交互:

$$\mathcal{F}_i = \underbrace{f_i \otimes f_i \otimes \cdots \otimes f_i}_{P\text{-order}} \tag{2}$$

其中, $\mathcal{F}$ 是指进行P阶外积之后的张量, $\otimes$是指张量直积. 经过直积运算之后虽然模态之间拥有足够的交互性, 但其维度过高, 且包含较多冗余信息, 直接放入协同过滤分解模型难以训练得到有效结果, 因此需要再通过一个池化层, 获得更为紧凑的融合特征. 最终的单分支融合特征可以表示为:

$$fusion\_f_i = W^H \cdot \mathcal{F}_i \tag{3}$$

此时, 式 (3) 中的W^H的维度呈指数级增长, 因此, 使用低秩张量网络进行降维, 采用基于秩-RCP 的张量分解方法:

$$\begin{aligned} fusion\_f_i &= \sum_{r=1}^{R} a_r^h \prod_{p=1}^{P} w_a^{h(p)} \prod_{p=1}^{P} f_{i(p)} \\ &= \sum_{r=1}^{R} a_r^h \prod_{p=1}^{P} w_a^{h(p)} f_{i(p)} \end{aligned} \tag{4}$$

其中, $\left\{\left\{a_r^h, w_a^h\right\}_{r=1}^{R}\right\}_{h=1}^{H}$可以看做是融合过程中的参数值, 使用张量分解能够显著降低全连接过程中的参数维度.

对于第 2 个阶段: 基本步骤与第 1 个阶段相同, 仅在第 1 个级联步骤需要将在 3 个分支的操作结果上进行, 即:

$$f_i' = Concat\left(M_x, M_y, M_z\right) \tag{5}$$

其中, M_x, M_y, M_z分别表示需要融合的 3 个分支上的模态特征. 其余步骤与第 1 阶段相同, 此处不再赘述. 该模块最终输出的融合特征为: $fusion\_f_i'$.

1.3 深度协同过滤模块

深度协同过滤模块的目的是将融合模态输入堆叠降噪自编码器 (SDAE) 模型中, 输出协同过滤分解的辅助信息以进行最终的推荐. 该模块通过二阶段张量融合后的特征向量表征物品, 能够更好地提升推荐系统的性能.

基于 SDAE 的协同过滤推荐算法[23] 与传统的基于矩阵分解的协同过滤推荐算法[24] 类似, 都是将评分矩阵 R 分解为用户潜在特征向量 u 和物品潜在特征向量 v 的方式. 基于此, 用户 m 对物品 n 的预估评分可以写为:

$$\hat{r}_{mn} = u_m^{\mathrm{T}} v_n \tag{6}$$

在 SDAE 模块, 本文所使用的 SDAE 模型由 5 层 DAE 组成, 前两层为编码器, 其映射函数如下:

$$H_i = dropout(ReLU\left(W_i H_i + b_i\right)) \tag{7}$$

其中, i 为 SDAE 的层编号, 第 1 层的H_{i-1}为加了噪声的融合特征$fusion_f_i'$.

后两层为解码器, 其映射函数如下:

$$Y_l = ReLU\left(W_l H_{l-1} + b_i\right) \tag{8}$$

同时, 需要从该模型中的第 2 层输出中获取物品融合特征的融合语义表达:

$$v' = dropout(ReLU\left(W_2 H_1 + b_2\right)) \tag{9}$$

1.4 目标函数

本文的目标函数使用联合损失函数, 联合损失的好处是可以通过多个目标对整个推荐系统的各个模块进行优化, 使其成为一个紧耦合的训练过程. 参考文献 [23] 中的联合损失函数, 为提高计算效率, 本文简化了其中 SDAE 每一层输出与原始输出的优化损失函数.

为防止模型过拟合, 需要在损失函数上增添规范化因子. 在 SDAE 模型中, 使用每一层权重和偏置的 L2 正则化表达; 在矩阵分解模型中, 使用分解后用户潜在向量的 L2 正则化表达:

$$L_{\mathrm{SDAE}} = \sum_l \left(\frac{1}{2} W_l^2 + \frac{1}{2} b_l^2\right) \tag{10}$$

$$L_{\mathrm{MF}} = \sum_m \|u_m\|^2 \tag{11}$$

$$L_1 = \lambda_u L_{\mathrm{MF}} + \lambda_w L_w \tag{12}$$

对于物品的潜在向量v的优化使用如下所示损失函数, 其目标为减少物品的融合特征矩阵 v 与评分矩阵 R 分解得来的v_n之间的误差, 使得物品特征中含有评分矩阵v_n的特征:

$$L_2 = \lambda_v \sum_n \left\|v_n - v'\right\|^2 \tag{13}$$

对于 SDAE 编码器去噪效果的优化, 其目标为减少输入的融合特征与重构特征的误差:

$$L_3 = \lambda_n \sum_n \left\|Y_L - fusion\_f_i'\right\|^2 \tag{14}$$

对于评分矩阵R分解过程的优化, 使用实际评分与预测评级之间的平方误差:

$$L_4 = \lambda_w \sum_m \sum_n (r_{mn} - \hat{r}_{mn})^2 \tag{15}$$

最终的损失为正则化因子与目标的损失函数之和:

$$L = L_1 + L_2 + L_3 + L_4 \tag{16}$$

1.5 行列式点过程 (DPP) 多样性推荐模块

该模块利用模态融合模块中的特征向量构建电影之间的相似度矩阵, 再结合用户对电影的评分构建用户的核矩阵, 使用贪心算法完成对核矩阵子式的快速求解, 最终输出用户的推荐列表. 旨在兼顾推荐准确性的同时, 实现对用户的多样性推荐.

行列式点过程 (determinantal point process) 是一个概率模型[25], 是一种用于多样性推荐的抽样方法. 点过程P是指集合$M=\{m_1,m_2,\cdots,m_n\}$所有子集 (2^M) 上的概率分布. 此时, 若存在半正定矩阵L确定点过程P, 且由M中元素索引, 那么称点过程P为行列式点过程, 矩阵L为核矩阵. 即当M为行列式点过程P所得的子集时, 对任意集合$A\subseteq M$, 有式 (17) 成立:

$$P(A\subseteq M)=det(L_A) \tag{17}$$

其中, $det(L_A)=\left[L_{i,j}\right]_{i,j\in L}$. 例如, 当集合 A 为电影 i 和电影 j 时:

$$\begin{aligned} det(L_A)&=\begin{vmatrix} L_{ii} & L_{ij} \\ L_{ji} & L_{jj} \end{vmatrix}=L_{ii}L_{jj}-L_{ij}L_{ji} \\ &=P(i\in M)P(j\in M)-L_{ij}^2 \end{aligned} \tag{18}$$

由式 (18) 可知, 核矩阵L的主对角线元素为某一部电影被选择到子集中的概率, 而非主对角线元素表示了两部电影之间的相似度. 因此, 两部电影相似度越高, 它们同时被选择的概率就越低. 行列式点过程通过计算核矩阵的子式, 在兼顾推荐准确性的同时, 巧妙地考虑了推荐的电影之间的多样性.

DPP 从待选电影集合中选择能最大化后验概率 (MAP) 的电影子集. 但是, DPP 的 MAP 直接求解比较复杂, 而贪心算法能大大加快求解速度[26]. 即每次选择一部电影 j 添加到结果集M_g中, M_g初始化为空集, 电影 j 满足式 (19):

$$j=\text{argmax}\ \log det\left(M_g\cup\{i\}\right)-\log det\left(M_g\right) \tag{19}$$

其中, 电影$i\in\dfrac{M}{M_g}$, 对L_{M_g}做柯列斯基分解可以简化行列式计算, 即存在非异下三角矩阵V, 使得$L_{M_g}=VV^{\mathrm{T}}$. 因此, $L_{M_g\cup\{i\}}$的柯列斯基分解如下:

$$L_{M_g\cup\{i\}}=\begin{bmatrix} L_{M_g} & L_{M_g,i} \\ L_{i,M_g} & L_{ii} \end{bmatrix}=\begin{bmatrix} V & 0 \\ c_i & d_i \end{bmatrix}\begin{bmatrix} V & 0 \\ c_i & d_i \end{bmatrix}^{\mathrm{T}} \tag{20}$$

由矩阵乘法, 行向量c_i和标量d_i满足:

$$Vc_i^{\mathrm{T}}=L_{M_g,i} \tag{21}$$

$$d_i^2=L_{ii}-c_{i2}^2 \tag{22}$$

此时, 求解式 (19) 归结于求解式 (23):

$$j=\text{argmax}\log\left(d_i^2\right) \tag{23}$$

式 (23) 求解后, 新的电影 j 被选择到M_g, 此时对剩下的每部电影 i, 结合式 (20)–式 (22), c_i和d_i都可以进行更新, 更新后分别定义为c_i'和d_i', 它们满足式 (24) 和式 (25):

$$c_i'=\begin{bmatrix} c_i & \dfrac{\left(L_{ji}-\langle c_j,c_i\rangle\right)}{d_j} \end{bmatrix}\triangleq\begin{bmatrix} c_i & e_i \end{bmatrix} \tag{24}$$

$$d_i'^2=L_{ii}-c_i'^2_2=d_i^2-e_i^2 \tag{25}$$

使用贪心算法完成行列式点过程的 MAP 求解, 算法的每一次迭代都会推荐一部新的电影, 我们只需要根据模型需要设置终止条件, 即可快速实现电影的多样性推荐.

2 实验

当前多模态推荐使用的公开数据集主要有 Amazon 商品数据集、Book-Crossing 数据集、MovieLens 数据集、KWAI 短视频数据集等. 亚马逊和 Book-Crossing 数据集包含的物品模态均为图像和音频, KWAI 中包含了快手平台提供的短视频以及相应的文字介绍, MovieLens 平台提供了电影的简介数据以及相应电影的 IMDB 网页链接, 可以由此获取到相应的电影预告片视频. 本文的实验使用 MovieLens 进行实验.

为了使音频信息能够在计算效率较高的同时拥有更为丰富的多元信息, 考虑到电影的预告片中拥有具有代表性的人物对话、环境声音、有些预告片甚至有相应的主题曲配乐, 因此本文选用电影预告片的音频作为原始信息. 我们选取电影海报作为推荐系统的图像模态输入, 与视频中的画面帧相比, 电影海报图像与输入的声音具有更好的模态互补性. 通过爬虫爬取了 IMDB 网站[27] 上相应电影的文字介绍、电影海报和预告片音频作为 3 个模态的信息, 构建起的最终数据集的统计数据如表 2 和表 3 所示.

表 2 MovieLens-100k 数据统计表

类别	数目
用户	943
电影	1 682
评分	100 000

表 3 MovieLens-100k 物品多模态数据统计表

数据类型	数据量
简介文本	1 540
海报图像	1 654
预告片音频	1 311

2.1 实验设计

本文的实验环境采用 Windows 10 操作系统, CPU 配置为 Core i7-5820K CPU@3.30 GHz, 内存为 32 GB; GPU 为 GTX TITAN X, 显存为 12 GB, 使用 Python 3.9 进行代码编写, 实验框架为基于 TensorFlow 2.6.0 的深度学习框架, 同时, 本文使用 Cornac[28] 作为多模态推荐系统的基准比较框架.

本文所使用的优化器为 Adam 优化器, 训练集、测试集与验证集的数据的比例为 8:1:1, 关于实验中使用到的其他超参数, 我们经过调试后对其进行设置, 如表 4 所示.

表 4 实验过程参数设置表

参数名称	参数值
Batch size	128
Learning rate	0.001
Dropout rate	0.1
Hidden dimensions	[64, 32, 128]
Out dimension	200
Rank	4
λ_u	1
λ_v	1
λ_w	0.1
λ_n	1

2.2 评测指标

本文的实验使用推荐系统两个阶段指标: 召回阶段使用归一化折现累计增益 (*NDCG*), 排序阶段使用 *AUC* 进行衡量, 能够相对全面真实地对模型进行评价. 对于这两个指标来说, 结果均是越大越好.

归一化折现累计增益 (*NDCG*) 用于衡量长度为 k 的推荐列表与理想排名的接近程度:

$$DCG@k = \sum_{i=1}^{k} \frac{rel(i)}{\log_2(i+1)} \tag{26}$$

$$NDCG@k = \frac{DCG}{IDCG} \tag{27}$$

其中, $rel(i)$表示第 i 个物品的相关性得分, 在本文中使用最终评分来计算相关性得分, 分母中$\log_2(i+1)$是为减弱排在后面的推荐结果对 *DCG* 大小的影响, *IDCG* 是指根据相关性得分从大到小排序, 找到理想条件下的排序情况.

AUC 能够反映模型的排序能力, 它反映的是一个相对性:

$$AUC(R)_n = \frac{1}{|R|(n-|R|)} \sum_{r \in R} \sum_{r' \in 1,\cdots,n \backslash R} \delta(r < r') \tag{28}$$

其中, R 为推荐的 item 集合, $\delta(x)$表示当 x 为 true 时, 它为 1; 反之为 0.

同时, 为了评测多样性推荐的效果, 本文使用 *ILAD* 和 *ILMD* 两个评测指标:

ILAD 表示所有用户推荐列表中电影之间的平均距离 (负相关性), 计算方法如式 (29) 所示:

$$ILAD = \underset{u \in U}{\text{mean}} \underset{i,j \in R_u, i \neq j}{\text{mean}} (1 - S_{ij}) \tag{29}$$

其中, U表示用户集合, R_u表示用户 u 的推荐列表, S_{ij}表示电影 i 和电影 j 的相似度.

ILMD 表示所有用户推荐列表中电影之间的平均最小距离, 计算方法如式 (30) 所示:

$$ILMD = \underset{u \in U}{\text{mean}} \underset{i,j \in R_u, i \neq j}{\min} (1 - S_{ij}) \tag{30}$$

2.3 对比算法

2.3.1 准确性的对比算法

1) PMF[29]: PMF 是一个仅使用用户评分的基于概率矩阵分解的算法, 通过已知评分矩阵, 使用最大后验概率和最大似然估计得到 U 和 V 的特征矩阵, 再用特征矩阵去预测评分矩阵中的未知值.

2) WBPR[30]: WBPR 是基于贝叶斯个性化排名推荐算法 (BPR) 的改进, 同样仅使用用户评分进行推荐, 基于全局为负样本添加权重, 相比原始的 BPR 推荐结果有了显著提高.

3) VBPR[31]: VBPR 是一个基于视觉模态和评分信息的多模态推荐算法, 对物品的潜在特征与原始物品特征进行联合之后进行推荐.

4) AMR[6]: AMR 是一个基于视觉模态与评分信息的多模态推荐算法, 通过对图像增加噪声来训练鲁棒性更强的多模态推荐模型.

5) fusionCDL: 本文模型.

2.3.2 多样性的对比算法

1) MMR[32]: MMR 是用于文本摘要及推荐系统的最大边缘相关算法. 基于物品对于用户的相关性及物品间的相似度对推荐物品进行重排, 旨在给用户推荐相关物品的同时, 保证推荐结果的多样性.

2) MSD[33]: MSD 是基于最大边缘算法 (MMR) 的改进, 同样使用物品对于用户的相关性及物品间的相似度, 构建新的目标函数以更加全面地衡量推荐列表的多样性.

2.4 实验结果与讨论

2.4.1 与其他算法的比较

表 5 展示了本文模型与其他基线模型在 MovieLens-

100k 上 *NDCG* 指标的实验结果, 实验表明, 与基准实验相比, 本文的 fusionCDL 模型在推荐列表长度不同时, 均拥有最好的实验结果. 在多样性整合推荐之前先生成了较大的推荐列表用于多样性选择, 将推荐列表长度设置为 300–500 进行初步的融合与推荐准确性衡量. 实验结果中, 随着 *k* 值的增大, *NDCG* 有缓慢上升, 这与现实情况相符. 同时, *k* 的取值不同时, 推荐结果在 *NDCG* 这一指标上在均有最好的结果, 证明了本文算法的有效性.

表 5 本文模型与对比算法实验结果表 (*NDCG*)

k	PMF	WBPR	VBPR	AMR	fusionCDL
300	0.2263	0.2611	0.3066	0.3148	**0.3818**
320	0.2278	0.2989	0.3122	0.3226	**0.3861**
340	0.2350	0.2996	0.3197	0.3321	**0.3912**
360	0.2406	0.3083	0.3165	0.3269	**0.3957**
380	0.2373	0.3041	0.3177	0.3380	**0.3998**
400	0.2431	0.3141	0.3240	0.3339	**0.4040**
420	0.2548	0.3150	0.3315	0.3371	**0.4070**
440	0.2524	0.3114	0.3277	0.3408	**0.4106**
460	0.2610	0.3183	0.3359	0.3443	**0.4132**
480	0.2574	0.3155	0.3340	0.3455	**0.4158**
500	0.2621	0.3176	0.3342	0.3475	**0.4186**

对初步推荐结果 *NDCG* 指标进行可视化如图 3 所示. 从图中可以明显看出, 在 *NDCG* 这一指标上, 之前的模型结果提升并不明显, 但 fusionCDL 相比之前的结果提升了 21.2% 左右, 效果十分显著, 说明模态融合能够大幅提升推荐算法的性能.

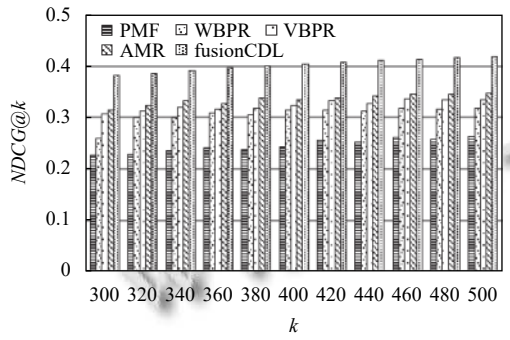

图 3 推荐列表长度 *k* 对 *NDCG* 的影响

表 6 展示了本文模型与其他基线模型在 MovieLens-100k 上 *AUC* 指标的实验结果, 本文在 *AUC* 这一指标上相对于 PMF 等模型拥有了显著的提高. 作为查全率和查准率的综合指标, *AUC* 结果超过 0.92, 且与融合了较好鲁棒性的多模态模型 AMR 相比, 实验结果也有相应提高, 证明了本文多模态融合推荐模型较好的综合性能.

表 6 本文模型与对比算法实验结果表 (*AUC*)

算法	*AUC*
PMF	0.7773
WBPR	0.8804
VBPR	0.9188
AMR	0.9238
fusionCDL	**0.9266**

在进行多样性的推荐时, 我们将最终推荐列表长度设置为 10, 从初步推荐列表的 500 个结果中使用 FGDPP 进行迭代, 以生成最终的多样性推荐结果, 实验结果如表 7 所示. 对比算法 MMR 与 MSD 给出的推荐电影列表间的平均距离 (*ILAD*) 与平均最小距离 (*ILMD*) 分别超过了 0.6 与 0.5, 这表明对比算法的实验结果已经具有较好的多样性. 在此基础上, 本文采用的 FGDPP 算法的实验结果在推荐多样性上有了进一步的提升, 在给用户推荐相关电影的基础上, 显著提高了电影之间的多样性.

表 7 多样性实验结果对比表

指标	MMR	MSD	FGDPP
ILAD	0.6228	0.6229	**0.7426**
ILMD	0.5334	0.5349	**0.6256**

2.4.2 消融实验

本文使用了文本、图像、音频 3 种模态的融合信息作为辅助信息, 为了更好地验证这 3 种模态信息以及融合模块对模型的贡献, 设计了如下消融实验.

首先在单模态和级联模态上进行了对比实验, 实验结果如表 8 所示. 可以看到, 对于级联方法, 将 3 种模态同时进行级联时 *AUC* 结果最好, 由此可以看出多模态特征对于推荐精排具有重要意义. 同时, 在单模态特征上, 图像模态作为辅助特征时实验结果最好, 说明视觉观感对于电影推荐拥有比文本、音频更为重要的意义. 除此以外, 在 *NDCG* 这一指标上, 单模态的结果比多模态结果要好, 分析原因是简单级联使得特征向量拥有较多冗余特征且没有充分考虑模态之间的交互, 从而使得 *NDCG* 结果有明显下降.

我们对模态融合进行研究, 旨在探究哪种模态对于融合结果影响最大, 以及三分支两阶段的融合方式是否能够提高模型性能. 对于两种模态的融合我们使用单分支单阶段的张量融合方法. 从表 8 中可以看出, 一方面, 模态融合之后, *NDCG* 结果有显著提升, 证明了融合模态对于推荐模型的有效性; 另一方面, 在几种

融合方式中, 3 种模态的三分支两阶段融合的 *NDCG* 与 *AUC* 均达到了最好的实验结果, 证明了本文特征融合算法的有效性. 同时, 在多种组合中, 融合了图像模态的推荐结果中, *NDCG* 和 *AUC* 指标均较高, 这也说明了视觉图像对于电影推荐重要意义, 与单模态和级联模态上的结果相符.

2.4.3 系统应用

本文构建了基于文本、图像、音频 3 种模态融合和基于 DPP 多样化结果精化的多模态推荐系统, 其特点如下: ① 输入 3 种模态数据, 系统的适用场景较为广泛; ② 使用张量融合方法生成物品的多模态特征向量, 能够提高推荐系统的推荐性能; ③ 进行多样性的结果精化, 从而提高推荐结果的多样性.

本文构建出了软件系统的应用模式, 如图 4 和图 5 所示, 系统的输入为当前数据库中物品的文本、图像和音频, 提取模态特征并与用户历史偏好信息一起输入本文模型中, 得到初步推荐结果后再进行多样性结果精化, 最后使用接口形式方便用户调用.

表 8 融合模态与单模态和级联模态的比较

级联方法			融合方法			*NDCG* @300	*AUC*
文字模态	图像模态	音频模态	文字模态	图像模态	音频模态		
√						0.3511	0.8926
	√					0.3524	0.9253
		√				0.3497	0.9215
√	√	√				0.3419	0.9262
			√		√	0.3809	0.9254
			√	√		0.3811	0.9255
				√	√	0.3802	0.9249
			√	√	√	**0.3818**	**0.9266**

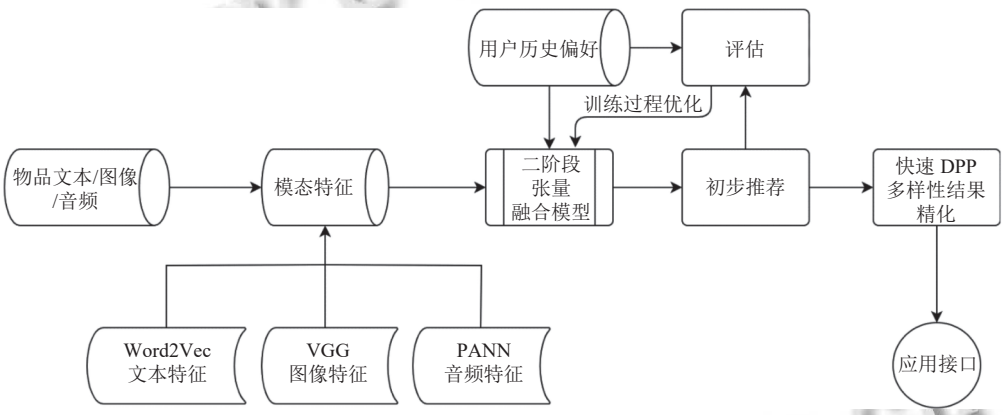

图 4 系统流程图

图 6 为本系统将 MovieLens 数据应用于电影平台的推荐样例, 本文为了使推荐结果更加清晰, 仅挑选一个用户推荐结果的图片以及电影标签进行展示, 可以看到, 上述电影虽然所属类别有共通之处, 但是各有不同, 能够充分满足用户在追求原有兴趣的同时探索多样结果的需求.

本文所提出的推荐系统, 不仅能够广泛应用于当前各类视频平台和音乐推荐中; 而且, 在各类线上展览、线上博物馆等平台更加注重沉浸式体验的当下, 使用展品图片、背景音频和文字介绍 3 种模态融合进行的推荐能够提升推荐系统的个性化程度, 从而带来更好的用户观览体验.

3 总结与讨论

本文使用基于张量网络的多模态算法, 使用文本、图像、音频 3 种模态数据经过两个阶段的张量网络融合生成融合特征, 在此融合特征的基础上进行深度协同过滤推荐, 之后再将推荐结果通过快速多样性算法将推荐结果进行多样化精化. 通过对比实验, 本文模型在 *NDCG* 这一指标上拥有显著提升, 在 *AUC* 这一指标上也具有最先进的实验结果, 证明了本文推荐模型较好的性能. 同时, 在多样性推荐方面上也实现了较好的结果, 实现了一个较为完备的推荐系统, 基于多模态数据的推荐系统具有较高的精确性和推荐结果多样性, 具有很好的实用性.

在改进与提升方面, 加入多样性推荐之后 *NDCG* 指标有相应下降, 分析原因是多样性训练过程并没有加入到整个模型的训练过程中. 即, 初次生成推荐列表之后加入多样性模块会对推荐有一定影响, 未来可以在整个模型的训练过程中进行多样性与准确性的综合考量.

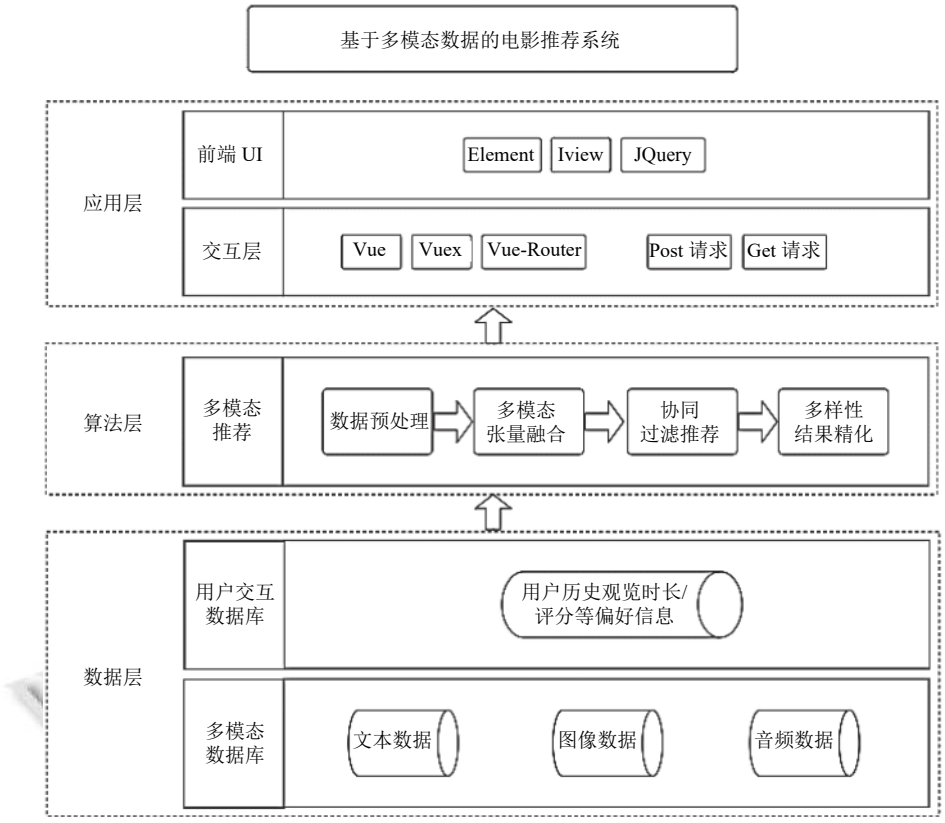

图 5 软件架构图

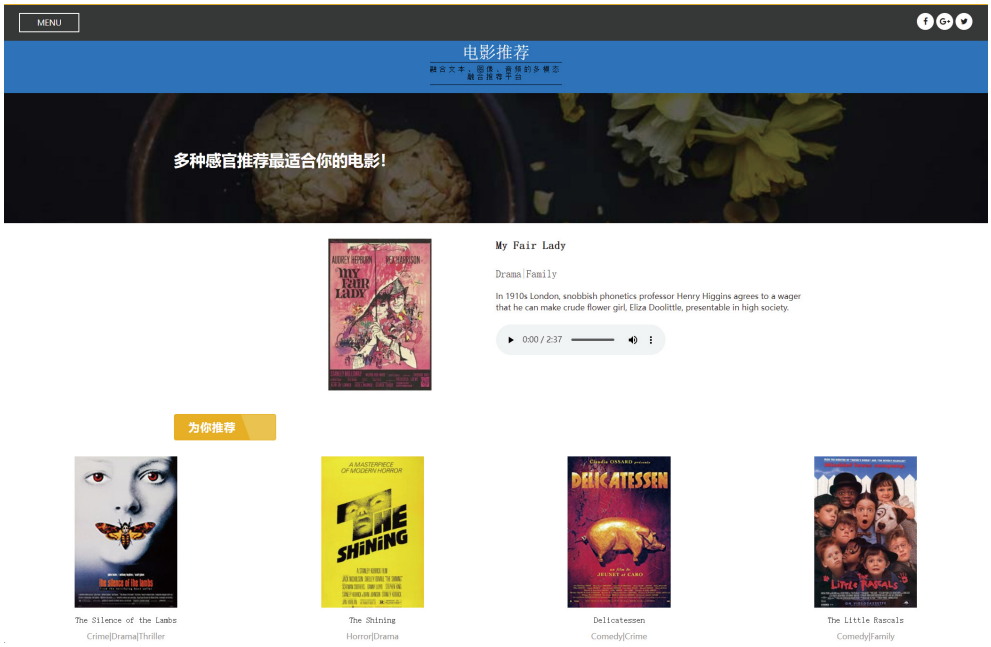

图 6 系统推荐样例

参考文献

1 Dong X, Yu L, Wu ZH, *et al*. A hybrid collaborative filtering model with deep structure for recommender systems. Proceedings of the 31st AAAI Conference on Artificial

Intelligence. San Francisco: AAAI Press, 2017. 1309–1315.
2 Truong QT, Salah A, Tran TB, *et al*. Exploring cross-modality utilization in recommender systems. IEEE Internet Computing, 2021, 25(4): 50–57. [doi: 10.1109/MIC.2021.3059027]
3 Wei J, He JH, Chen K, *et al*. Collaborative filtering and deep learning based recommendation system for cold start items. Expert Systems with Applications, 2017, 69: 29–39. [doi: 10.1016/j.eswa.2016.09.040]
4 Frolov E, Oseledets I. HybridSVD: When collaborative information is not enough. Proceedings of the 13th ACM Conference on Recommender Systems. Copenhagen: ACM, 2019. 331–339.
5 Lei CY, Li D, Li WP, *et al*. Comparative deep learning of hybrid representations for image recommendations. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vegas: IEEE, 2016. 2545–2553.
6 Tang JH, Du XY, He XN, *et al*. Adversarial training towards robust multimedia recommender system. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(5): 855–867. [doi: 10.1109/TKDE.2019.2893638]
7 Qiu RH, Wang S, Chen Z, *et al*. CausalRec: Causal inference for visual debiasing in visually-aware recommendation. Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021. 3844–3852.
8 Oramas S, Nieto O, Sordo M, *et al*. A deep multimodal approach for cold-start music recommendation. Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems. Como: ACM, 2017. 32–37.
9 Sun MX, Li F, Zhang J. A multi-modality deep network for cold-start recommendation. Big Data and Cognitive Computing, 2018, 2(1): 7. [doi: 10.3390/bdcc2010007]
10 肖庆华, 刘学军, 施浩杰. 多模态下的互补物品的多样性推荐. 小型微型计算机系统, 2021, 42(9): 1859–1864. [doi: 10.3969/j.issn.1000-1220.2021.09.010]
11 Huang PY, Patrick M, Hu JJ, *et al*. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021. 2443–2459.
12 Wang SN, Zhang JJ, Zong CQ. Associative multichannel autoencoder for multimodal word representation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 115–124.
13 Zadeh A, Liang PP, Morency LP. Foundations of multimodal Co-learning. Information Fusion, 2020, 64: 188–193. [doi: 10.1016/j.inffus.2020.06.001]
14 Liu J, Zhu XX, Liu F, *et al*. OPT: Omni-perception pre-trainer for cross-modal understanding and generation. arXiv: 2107.00249, 2021.
15 Hou M, Tang JJ, Zhang JH, *et al*. Deep multimodal multilinear fusion with high-order polynomial pooling. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 12156–12166.
16 梁美玉, 王笑笑, 杜军平. 基于多模态图和对抗哈希注意力网络的跨媒体细粒度表示学习. 模式识别与人工智能, 2022, 35(3): 195–206. [doi: 10.16451/j.cnki.issn1003-6059.202203001]
17 Lin TY, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1449–1457.
18 Zadeh A, Chen MH, Poria S, *et al*. Tensor fusion network for multimodal sentiment analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, 2017. 1103–1114.
19 Al-Assam S, Clark SR, Jaksch D. The tensor network theory library. Journal of Statistical Mechanics: Theory and Experiment, 2017, 2017: 093102. [doi: 10.1088/1742-5468/aa7df3]
20 徐俊, 张政, 杜宣萱, 等. 基于项目语义的协同过滤冷启动推荐算法研究. 小型微型计算机系统, 2021, 42(11): 2246–2251. [doi: 10.3969/j.issn.1000-1220.2021.11.002]
21 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.
22 Kong QQ, Cao Y, Iqbal T, *et al*. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880–2894. [doi: 10.1109/TASLP.2020.3030497]
23 Wang H, Wang NY, Yeung DY. Collaborative deep learning for recommender systems. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM, 2015. 1235–1244.

24 Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer, 2009, 42(8): 30–37. [doi: 10.1109/MC.2009.263]

25 Kulesza A, Taskar B. Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning, 2012, 5(2–3): 123–286.

26 Chen LM, Zhang GX, Zhou HN. Fast Greedy MAP inference for determinantal point process to improve recommendation diversity. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 5627–5638.

27 IMDb: Ratings, reviews, and where to watch the best movies & TV shows. https://mungfali.com/post/2B6635BC5C356BA561ED0C20A770FA2019262FA4. [2022-04-10].

28 Salah A, Truong QT, Lauw HW. Cornac: A comparative framework for multimodal recommender systems. Journal of Machine Learning Research, 2020, 21(95): 1–5.

29 Salakhutdinov RR, Mnih A. Probabilistic matrix factorization. Proceedings of the 20th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2007. 1257–1264.

30 Gantner Z, Drumond L, Freudenthaler C, *et al*. Personalized ranking for non-uniformly sampled items. Journal of Machine Learning Research, 2012, 18: 231–247.

31 He RN, McAuley J. VBPR: Visual Bayesian personalized ranking from implicit feedback. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press, 2016. 144–150.

32 Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne: ACM, 1998. 335–336.

33 Borodin A, Lee HC, Ye YL. Max-sum diversification, monotone submodular functions and dynamic updates. Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. Scottsdale: ACM, 2012. 155–166.

(校对责编: 孙君艳)