本科生毕业设计（论文）开题报告

题　　目：基于多模态表征学习的序列推荐算法

院　　系_____计算机科学与技术学院_____

专业班级_____CS1806_____

姓　　名_____刘美_____

学　　号_____U201814788_____

指导教师_____李剑军_____

2022 年 3 月

一、课题研究背景

在互联网信息高速发展的今天，推荐系统发挥着非常重要的作用。它能够帮助用户在使用 Web 应用时，从大量的信息中筛选出其感兴趣的信息，从而缓解信息过载问题。目前现有的很多推荐算法都是基于用户信息及其交互行为会被记录且能够因此判定用户的偏好研究的，然而事实是，很多用户的行为是隐式的，我们难以从这些行为中获取到足够的有效信息，比如用户点击某个商品，并不代表其厌恶/喜爱这个商品。除此以外，用户的信息也不一定总能获取，例如当其以游客方式访问或者由于隐私保护政策时。因此大多数情况下，我们实际能够获取到的数据是稀疏、难以产生准确结果的。

但其实，现实生活中的多媒体数据往往是多种信息的传递媒介（例如，一段视频中往往会同时有文字信息、视觉信息和听觉信息的传播）。因此，可以将多模态信息引入到推荐系统中，使模型在训练过程中得到更好获得用户和物品表示，能够有效地缓解用户-物品交互的数据稀疏性问题以及数据冷启动发问题。

另一方面，用户和商品的交互本质上是序列依赖的。现实生活中，用户的购物习惯通常是一整个序列化的行为而不是孤立发生的。例如，购买了咖啡机后，就会购买咖啡豆，后面的行为取决于先前的动作。序列化的依赖关系通常存在于交易数据中，但常规的基于内容和基于协同过滤的推荐系统都很难捕获这种序列依赖关系。这从根本上促进了序列推荐系统的发展。

同时，用户的偏好和物品的受欢迎程度也是随着时间变化而变化的，不是静态的，用户对物品偏好的动态演变以及物品本身的价值变化，对提高推荐系统的预测准确性有很大的意义，并且这种动态变化只能由序列推荐算法捕获。

序列推荐算法将用户和物品地交互建模为一个动态的序列，并且利用序列地依赖性来捕获用户当前以及最近的偏好，实现更加准确和动态的推荐。

二、国内外研究现状

2.1 多模态表示学习

多模态表示学习的目的是将被研究对象（结构化数据、图像、视频、文本、语音等）中蕴含的语义信息抽象为实值向量。当多个模态共存时，我们需要同时从多个异质信息源中提取被研究对象的特征。在单模态表示学习的基础上，多模态表示学习还要考虑多个模态信息的一致性和互补性。目前多模态表示学习分为

两类：联合表示和协同表示。

联合表示是将多个模态的信息一起映射到一个统一的多模态向量空间。联合表示最简单的示例是单个模态特征的串联（也称早期融合[1]）。其次，还有基于神经网络、图像模型的方法。基于神经网络的联合表示通常具有优越的性能，并能在无监督的情况下对表示进行预训练[2]。然而，其性能的提高取决于训练的数据量。其次，模型不能自然地处理丢失数据，尽管目前存在一些方法可以缓解这个问题[3]，但仍然是有限的。基于图像模型的方法是一种通过潜在随机变量来构造表示的常用方法[5]。目前这方面最流行的方法是受限玻尔兹曼机（DBM）[6]。使用多模态 DBMs 学习多模态表示的最大优点之一是其生成特性允许以一种简单的方式处理丢失的数据——即使整个模态丢失，模型也有一种自然的处理方法。它还可以用于在另一种模态存在的情况下生成一种模态的样本，或从表示中生成两种模态的样本。与自动编码器类似，可以以非监督的方式对表示进行预训练，从而支持使用未标记的数据。DBMs 的主要缺点是训练困难，计算成本高，需要使用近似变分训练方法。

协同表示则将多模态中的每个模态分别映射到各自的表示空间，但映射后的向量之间满足一定的相关性约束（如线性约束）。早期的相似模型[7]通过最小化协同空间中模态之间的距离来进行约束，例如维斯顿等人的研究。在 WSABIE 中，为图像及其注释构建了一个协同空间。其从图像和文本特征中构造了一个简单的线性映射，这样相关的注释和图像表示在它们之间会比不相关的注释与图像表示有更高的内积。相似模型强调表示之间的相似性，但结构化的协同表示空间模型超越了这一点，并且在模态表示之间强制执行额外的约束，具体的结构化限制根据应用而异。结构化协同通常用于跨模式哈希中，将高维数据压缩为紧凑的二进制代码，并对类似对象使用相似的二进制代码[18]。哈希的方法最终迫使多模态空间表示有一些限制：（1）不同模态的相同对象有相似的哈希编码；（2）多模态空间必须保持数据相似性。

联合表示将多模态数据投射到一个公共空间中，最适合在推理过程中出现所有模态的情况。另一方面，协同表示法将每个模态投影到一个单独但协同的空间中，使其适用于测试时只有一个模态的应用。

2.2 序列推荐算法

2.2.1 研究成果

从技术角度来说，目前基于序列算法的推荐系统主要分为三个大类：传统序列模型，潜在表征模型和深度神经网络模型。

传统序列模型包括序列模式挖掘[8]和 Markov Chains 模型[9]等。这种方式通过利用序列推荐系统的自然优势（用户-商品的交互本质上存在顺序依赖性）在一个序列中对用户-商品的交互上建立顺序依赖关系。

潜在表征模型首先学习一个用户或商品的潜在表征，然后利用这个学习到的潜在表征去预测后续的用户-商品交互。因此，更多隐形和复杂的依赖关系在一个潜在空间中被捕获，极大地改善了推荐系统的推荐性能。

最近几年，深度神经网络由于其对于在一个序列中不同实体间构建和捕获综合关系具有自然的优势，几乎主导了序列推荐系统，包括了如基于图神经网络的推荐算法[11]、基于卷积神经网络的推荐算法[12]以及基于循环神经网络[14][14]等基础深度神经网络序列推荐系统，以及如引入注意力模型、记忆网络和混合模型等的高级深度神经网络序列推荐系统。

2.2.2 挑战与不足

由于现实世界用户行为、物品特征、交互环境的多样性，生成的用户-物品数据往往具有不同的特征，因此为序列推荐算法带来了不同的挑战。

1. 难以学习到高阶依赖关系。高阶顺序依赖关系由于是跨多个用户项交互的复杂多级级联依赖关系，因此很难被捕获。目前解决高阶依赖关系建模难的主要有两种方式：（1）基于高阶马尔可夫链方法[15]；（2）基于 RNN 的方法[16]。但是这两种方法也才存在诸多局限性，例如，随着模型参数的数量随阶数呈指数增长，高阶马尔科夫链模型中可能涉及到的历史状态非常有限，而 RNN 中采用的过强假设限制了 RNN 再序列中的灵活应用。
2. 难以学习长期依赖关系。目前解决长期依赖问题，一般是引入 LSTM 或基于 GRU 的模型，但是 RNN 模型高度假设相邻的物品之间存在高度的相关性，导致容易产生错误的依赖。目前还有一种混合的模型[17]，将多个子模型与不同时间反俄文欸的子模型结合起来，在一个统一的模型中

捕获长期和短期的依赖关系，但这种方法仍然相当有限。

3. 用户-物品交互的顺序是灵活的。所谓灵活是指，用户-物品的交互时而有序，时而无序。对于这种灵活的顺序，最好捕获集合顺序依赖关系，而不是点式依赖关系，因为前者是模糊的，并且不假设用户项交互的严格顺序。现有的基于马尔科夫链、因式分解机或 RNN 的序列推荐模型都是解决点式依赖的，并不能很好地解决集合依赖顺序。

三、课题研究的意义、内容和目标

3.1 课题研究的意义

现实生活中，用户的购物习惯往往是一整个序列化行为，而不是独立的，且用户的偏好会随着时间的推移而改变。序列推荐系统(SRS)试图理解和建模用户的连续行为以及用户偏好随时间的变化，从而实现动态地捕获用户的行为演化和偏好。同时通过利用多模态信息能够给序列推荐系统提供丰富的特征和信息，有效地缓解数据稀疏问题。将多模态实体特征与序列特征相结合进行物品推荐，有助于提高推荐的准确性和可解释性。

3.2 课题研究的内容

1. 利用不同的方法对不同模态数据学习不同的嵌入表征；
2. 设计序列推荐模型，将多模态实体特征与序列特征相结合进行物品推荐；
3. 设计演示系统，对设计实现的基于多模态知识图的序列推荐算法进行效果演示。

3.3 课题研究的目标

1. 提出的基于多模态知识图的序列推荐算法在相关性能指标上有一定的提升；
2. 演示系统能够有效演示。

四、技术关键与技术路线

4.1 多模态表征学习

多模态表征学习，主要包括图像表征、文本表征和属性表征三个方面的学习，是为了提取不同模态的特征信息。提取的过程如图所示。

属性表征：对于 item 的各个属性，如 id、类别、颜色等信息进行编码。对于这些信息，使用 One-hot 方法编码的向量会很高维也很稀疏，因此我们使用

Embedding Layer 的方法，首先获得属性信息的 One-hot 矩阵作为输入，再通过一个全连接神经网络层，将不同信息映射到致密的低维向量中。Embedding Layer 的输出维度 dx 将作为可选择的参数。

文本表征：使用 Word2Vec[18]来训练词向量，然后使用 SIF[19]模型，将预训练好的词向量，然后使用加权平均的方法，对句子中所有词对应的词向量进行计算，得到整个句子的 embedding 向量，最后使用主成分分析去掉一些 special direction，即在完成词加权平均后，移出所有行为向量的公共主成分，只保留反应序列特性部分，来表示文本特征。

图像表征：使用 ResNet50[20]的最后一个隐藏层的 2048 维特征（由 Inmagenet[21]训练得到）进行训练获得。

4.2 多模态序列图推荐

由于不同模态之间存在语义差异，且不同用户可能对不同模态偏好程度不同，因此为每个模态单独建立一个序列图，在每个序列图上执行相同的操作。某个模态得到序列嵌入和节点嵌入的工作流程如图 4.1 所示。

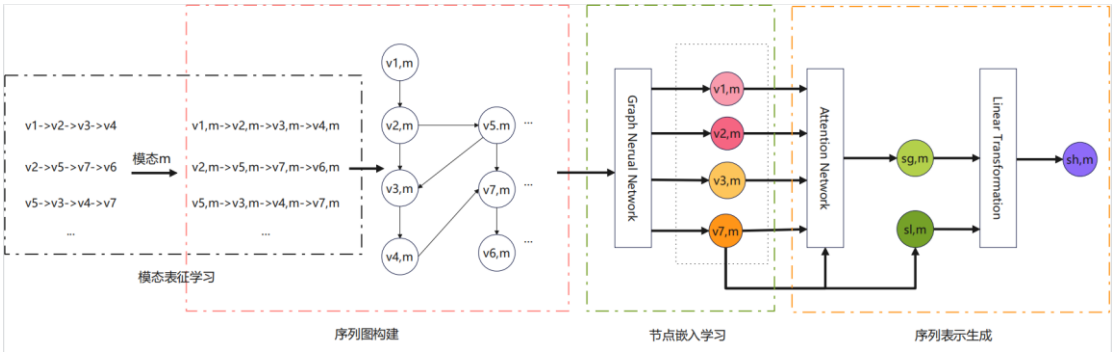

图 4.1 学习序列嵌入与节点嵌入流程图

4.2.1 原始序列图构建

对于某个模态，所有用户的所有序列被建模为一个有向序列图，其中每个用户的序列被视为一个子图。对于某个用户序列子图表示为：$G_m = (\mathrm{V}, \mathrm{E})$。图中每个节点$\mathrm{v} \in \mathrm{V}$代表用户交互的某个节点 i，其表示为上游通过多模态表征学习获得。每条边$(v_{i-1}, v_i) \in E$意味着用户与物品 i-1 交互后又与物品 i 进行交互。由于在一个序列中，用户可能多次与同一个物品进行交互，因此，对每条边分配一个归一化的权重，该权重的计算方式是边的出现次数除以这条边源节点的出度。

由于在多模态表征学习中到的不同节点的维度可能是不一样的，我们需要将所有节点统一到同一个维度空间。节点最后统一维度的表示将通过 GNN 网络获得，对应的序列因此可以被表示为由节点张量组成的嵌入张量s_m。

4.2.2 序列图中节点的嵌入学习

我们通过图神经网络获得节点的隐藏张量表示。对于子图G_m中的节点v_i来说，更新函数表示如下：

$$a_i^t = A_i[v_1^{t-1}, \dots, v_n^{t-1}]^T H + b, \tag{1}$$

$$z_i^t = \sigma(W_z a_i^t + U_z v_i^{t-1}), \tag{2}$$

$$r_i^t = \sigma(W_r a_i^t + U_r v_i^{t-1}), \tag{3}$$

$$\widetilde{v_i^t} = \tanh\left(W_o a_i^t + U_o(r_i^t \odot v_i^{t-1})\right), \tag{4}$$

$$v_i^t = (1 - z_i^t) \odot v_i^{t-1} + z_i^t \odot \widetilde{v_i^t}, \tag{5}$$

其中，$H \in R^{d \times 2d}$控制权重，$z_{s,i}, r_{s,i}$分别代表重置门和更新门。$[v_1^{t-1}, \dots, v_n^{t-1}]$是序列$s_m$中的节点列表，$\sigma(\cdot)$是 sigmoid 函数，$\odot$是元素乘法操作。$v_i \epsilon R^d$是节点$v_i$的潜在表示。关系矩阵$A_s \epsilon R^{n \times 2n}$决定了图中节点之间是如何交流的，$A_{s,i} \epsilon R^{1 \times 2n}$则表示$A_s$中和节点$v_i$对应的两列。

对于序列图G_m，门图神经网络同时处理所有节点。在矩阵A_s给出的前提下，公式（1）用来实现不同节点之间的信息传播。它提取邻居的潜在表示，并将信息输送到图神经网络的输入，然后利用更新和重置门，决定哪些被传播的信息被保留，哪些被忽略。再之后，按照公式（4），根据节点的历史表示，重置门以及当前状态构建候选的节点表示。最后的状态是在更新门的控制下，由历史隐藏状态和候选状态结合而成。直到更新完序列图中的所有节点收敛，我们才得到最后的节点潜在表示。

4.2.3 序列的嵌入学习

序列的表示是由序列中的节点直接表示而来的。为了提高推荐的准确性，我们将长期的偏好与序列中最近时间的兴趣结合，使用混合的嵌入作为序列的嵌入。

对于序列$s_m = [v_1, \dots, v_n]$，其局部嵌入可以简单表示为最后一次交互的节点的嵌入，即s_l=v_n。其全局嵌入，可以通过序列图中所有节点的表示聚合而成。我们使用 soft-attention 机制获取全局嵌入s_g的表示，然后，将局部嵌入s_l与全局嵌

入s_g的拼接通过线性转换获得最终的序列嵌入表示。

4.2.4 模态的融合

分别在三个模态序列子图获得序列的最终表示$s_{h,1}$，$s_{h,2}$和$s_{h,3}$之后，我们探索使用两种不同的方法进行模态融合对最后推荐的性能的影响：

1. 对不同模态的序列嵌入与节点嵌入进行串联拼接：使用非线性变换，将三个表征拼接起来，对应的，对三个模态的每个节点执行同样的操作拼接得到每个节点三个模态的融合表示。

$$s_h = LeakyReLu\left(W_4\left(s_{h,1} \| s_{h,2} \| s_{h,3}\right)\right), \tag{6}$$

$$v_i = LeakyReLu\left(W_5\left(v_{i,1} \| v_{i,2} \| v_{i,3}\right)\right), \tag{7}$$

2. 对不同模态得序列嵌入与节点嵌入进行线性变换。下面展示了三个模态的序列嵌入线性变换实现模态融合的公式。序列嵌入同理。

$$s_h = LeakyReLu\left(W_6 s_{h,1} + W_7 s_{h,2} + W_8 s_{h,3} + b\right), \tag{8}$$

4.2.5 预测推荐

将最后的多模态序列融合表示s_h与每一个候选的 item 的多模态融合表示v_i，通过计算得到其得分函数$\hat{z}_i$，然后对所有 item 的得分函数进行拼接得到最后的得分矩阵$\hat{z}$。然后将得分矩阵$\hat{z}$通过一个 softmax 函数，得到预测的用户下一个交互的物品为真正交互的物品的概率。概率越大，用户越有可能进行交互。

对于每个序列图，损失函数为预测结果与实际结果的交叉熵。

五、风险分析

风险来源一：可能欠缺研究该课题的知识或技能，主要表现在部分未经过系统学习的科目。比如机器学习、深度学习等。本课题涉及到对各个模态进行不同的嵌入处理，要求熟练应用常用的深度学习模型和自然语言处理模型。除此之外，可能会有其他更多的需要补充的知识。就对应措施而言，如果出现了必须要进行大量学习的知识漏洞，应当及时入手相关教材，并主动学习所欠缺的知识，必要时会与导师进行商讨。

六、课题研究进度安排

表 6.1 课题研究进度安排表

学期	周次	工作任务
2021-2022 第一学期	第 8 周	联系导师、商报课题并申报、确认
	第 9 周——第 14 周	阅读课题相关文献与实现源码；学习课题涉及的基础知识；对所选课题方向进行调研
	第 15 周	确定课题研究算法的初步实现路线
	第 16 周——第 19 周	商讨并确定算法的关键技术；开始设计代码
2022-2022 第二学期	第 2 周	完成文献翻译；撰写开题报告；准备开题答辩演示文件；进行开题答辩
	第 8 周	完成毕设中期检查
	第 9 周——第 12 周	完成论文
	第 12 周——第 14 周	论文形式检查；论文查重
	第 16 周	毕业答辩

七、主要参考文献

[1] M. Denkowski and A. Lavie, "Meteoruniversal: Language specific translationevaluation for any target language,"in Proc.14th Conf. Eur. Chapter Assoc. Compute. Linguistics, 2014.

[2] A. Haubold and J.R. Kender, "Alignment ofspeech to highly imperfect text transcriptionin Proc.IEEE Int. Conf. Multimedia Expo, 2007, pp. 224-227

[3] M. Mller, "Dynamic time warping,"in Inform. Retrieval for Music and Motion. Berlin, Germany: Springer, 2007, pp. 69-84

[4] M. Valstar, et al., "AVEC 2013, ”in Proc. ACM Int. Workshop Audio/Visual Emotion Challenge, 2013, pp.3-10.

[5] L.W. Barsalou,"Grounded cognition,”Antnu. Rev. Psychology, vol.59, pp. 617-645,2008.

[6] N. Rasiwasia, et al., "A new approach to cross-modal multimedia retrieval," in Proc.19th ACM Int. Conf. Multimedia, 2010, pp. 251-26

[7] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in Proc. 19th ACM Int. Conf. Multimedia, 2014, pp. 7-16.

[8] Shanshan Feng, Xutao Li, Yifeng Zeng, Gao Cong, and Yeow Meng Chee. Personalized ranking metric embedding for next new poi recommendation. In Proceeding soft the 24h International Join tConference on Artificial Intelligence, pages 2069- -2075, 2015.

[9] Ghim-EngYap, Xiao-LiLi, and PhilipYu. Effective next-items recommendation via personalized sequential pattern mining. In Database Systems for Advanced Applications, pages 48- -64, 2012

[10] Florent Garcin, Christos Dimitrakakis, Boi Faltings. Personalized news recommendation with context trees. Proceedings of the 7th ACM conference on Recommender systems October 2013 Pages 105-112

[11] Shu Wu, Yuyuan Tang, and et al. Sessionbased recommendation with graph neural networks. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, pages 1-9, 2019.

[12] Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In Proceedings of the 11th ACM International Conference on Web Search and Data Mining, pages 565- 573, 2018.

[13] Fajie Yuan, Alexandros Karatzoglou, loannis Arapakis, Joemon M Jose, and Xiangnan He. A simple convolutional generative network for next item recommendation. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, pages 582- -590, 2019.

[14] Massimo Quadrana, Alexandros Karatzoglou, and et al. Personalizing session-based recommendations with hierarchical recurrent neural networks. In Proceedings of the 11th ACM Conference on Recommender Systems, pages 130-137, 2017.

[15] Ruining He, Julian McAuley. Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation.

[16] Balazs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, Domonkos Tikk. Session-based Recommendations with Recurrent Neural Networks.

[17] Jiaxi Tang, Francois Belletti, Sagar Jain, Minmin Chen, Alex Beutel, Can Xu, Ed H. Chi. Towards Neural Mixture Recommender for Long Range Dependent User Sequences. The World Wide Web Conference May 2019 Pages 1782- -1793.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. 3111–3119.

[19] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. (2016).

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for

image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. Ieee, 248–255.

[22] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, Yong Li. Sequential Recommendation with Graph Neural Networks. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval July 2021 Pages 378- -387

[23] Rui Sun, Xuezhi Cao, Yan Zhao,JunchenWan, Kun Zhou, Fuzheng Zhang, ZhongyuaWang, Kai Zheng. Multimodal Knowledge Graphs for Recommender Systems. CIlKM 20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management October 2020 Pages 1405-1414

[24] Yinwei Wei, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, Kai Zheng. MMGCN: Multimodal Graph Convolution Network for Personalized Recommendation. CIKM 2020,1405-1414

[25] 周志华. 机器学习[M].北京:清华大学出版社, 2016 年 1 月

华中科技大学本科生毕业设计（论文）开题报告评审表

<table>
<tr><td>姓名</td><td>刘美</td><td>学号</td><td>U201814788</td><td>指导教师</td><td>李剑军</td></tr>
<tr><td colspan="2">院（系）专业</td><td colspan="4">计算机科学与技术</td></tr>
<tr><td colspan="6">指导教师评语
1. 学生前期表现情况。
2. 是否具备开始设计（论文）条件？是否同意开始设计（论文）？
3. 不足及建议。</td></tr>
<tr><td colspan="6">指导教师（签名）：
年　　月　　日</td></tr>
<tr><td colspan="6">教研室（系、所）或开题报告答辩小组审核意见</td></tr>
<tr><td colspan="6">教研室（系、所）或开题报告答辩小组负责人（签名）：
年　　月　　日</td></tr>
</table>