目 录

摘 要

近年来，图神经网络在越来越多的领域展现出重要作用，其中也包括推荐系统这一重要的互联网应用服务。目前，基于图神经网络的推荐系统存在以下问题：首先，推荐系统往往面临复杂的业务场景。不同场景的实体、实体之间的交互行为例如用户行为数据均不尽相同。这使得一般的图结构无法表达这些对象之间复杂的拓扑关系；其次，某些用户行为数据的量级往往对于交互对象的数量呈现稀疏的特征，使得一般的图表示学习方法很难在整个图上进行有效计算。因此，本文提出了基于超图注意力机制的图神经网络点击率预测模型 **HGAT-GNN**（**H**yper**G**raph **AT**tention **G**raph **N**eural **N**etworks）。其中，超图节点能够表示用户和物品等多类实体，而超图中的超边能够连接图中多个节点，均便于对复杂关系的表达。模型利用注意力机制从相邻超边中聚集和迭代节点表示，以用于后续的预测网络中。为应对用户行为数据稀疏的问题，迭代中采用的动态特征能够让节点特征的变化及时反映到图节点表示中，同时也可以用于减少图神经网络的参数量和计算量，提高模型整体的泛化性。本文将提出的模型运用于视频网站 bilibili 的两个真实数据集中。实验结果表明，模型的表现超过目前已知的其他相关最新方法，充分展现了模型的有效性。**HGAT-GNN** 模型的源代码可在 GitHub 代码仓库中查看[1]。

关键词: 图神经网络；超图；推荐系统；点击率预测；注意力机制

[1]https://github.com/codeplay0314/GNN-model

Abstract

In recent years, graph neural networks have shown increasing importance in more and more fields, including recommender systems, a typical Internet application. At present, several problems exist in the recommender systems based on graph neural networks: first, recommender systems often deal with complex business scenarios, and entities and the interaction between them, user behavior data for instance, vary in different scenarios, which makes the normal graph structure incapable of representing these sophisticated topological relationships; second, user behavior data are often sparse compared to the related entities, which can make the graph representation learning inefficient. In regard to these, this paper proposes a click-through rate prediction model **HGAT-GNN** (**H**yper**G**raph **AT**tention **G**raph **N**eural **N**etworks), based on the attention mechanism in the hypergraph neural networks. The nodes in the hypergraph represent entities like users or items, and hyperedges connect multiple nodes to facilitate the expression of complex relationships. The model utilizes attention mechanism to aggregate and update node representations from adjacent hyperedges for subsequent prediction networks. In order to deal with the problem of sparse user behavior data, this model uses the dynamic features in the updates of node representations to reflect the ever-changing nature of node features, and to reduce the amount of parameters and computation complexity of the model, which significantly improves the overall generalization of the model. In the experiment, the proposed model is applied to two real world datasets of the video website bilibili. The results show that the model outperforms other related state-of-the-art methods, fully demonstrating the effectiveness of the model. The source code of the **HGAT-GNN** model is released at [1].

Keywords: Graph neural networks; Hypergraph; Recommender systems; Click-through-rate prediction; Attention mechanism

[1] https://github.com/codeplay0314/GNN-model

第 1 章 绪 论

本章分为四个小节：论文背景一节介绍了本文的研究背景，从推荐系统的发展和图神经网络在此上的应用现状等方面探讨了基于超图的图神经网络推荐系统的可行性和必要性；研究难点一节探讨了常规推荐场景和用户行为数据稀疏的特殊场景下几点可能会面临的问题；论文的内容和意义一节总结了本文的工作和所做的几点贡献；章节结构一节阐述了本文之后的章节安排和内容。

1.1 论文背景

推荐系统近年来在互联网应用中越来越展现出其重要性。本世纪以来，互联网服务多以“门户网站”或“搜索”等形式展现，前者为所有用户提供了统一的信息界面，而后者则是用户通过自身需要主动对信息进行检索。随着互联网、移动手机、社交媒体以及大数据、数据挖掘、移动计算、云计算等计算机科学相关技术的发展，之前提到的这两种信息供应方式表现出效率不高、信息单一等一系列问题，而推荐系统便在这样的背景下应运而生。推荐系统能做到“千人千面”，给用户个性化地提供所需要的内容，极大提高了信息分发的效率。 图 1-1 展现了一个典型的互联网网站的推荐页面，它通过给用户提供与正在消费的物品相关的物品的展示，来吸引用户进行后续的消费，从而提高产品的效益。而与推荐系统相关的研究，近年来也越来越受到学术界和工业界的关注，有关推荐系统的目标、模型结构的设计、各阶段的策略、计算效率和实际场景中的问题都存在尚待解决的问题，是当下研究的一个热点课题。

推荐系统也随着人工智能技术的发展日新月异。以深度学习为代表的许多最新的机器学习技术的相关研究成果都能在非常快的时间内落地到推荐系统的应用上[1-3]。一个典型的例子是由谷歌公司提出的 Wide&Deep 点击率预测模型[4]，它利用深度神经网络的设计平衡了推荐模型的记忆能力和泛化能力，之后被广泛应用于众多场景中。近年来，随着对推荐系统的准确性和有效性要求的不断提高，点击率预测模型不断往着复杂和精细的方向发展。点击率预测模型利用给定的信息（特征），对用户在用户界面出现点击行为的概率进行预估，以最大化产品对于用户点击行为的吸引力。点击率预测模型以往通常会利用与用户和物品直接相关的特征，近年来，用户的行为数据也被视为能反映用户兴趣的重要来源，被用来提取用于推荐的关键信息。

用户行为数据因存在发生时间先后关系，可以被视作有着时间偏序关系的序列。众多将序列模型与推荐相结合的模型被提出后，被应用在实际场景中取得了较好的效果[5-6]。而随着近年来图神经网络研究的深入，图神经网络的架构也逐渐出现在推荐系统的应用中[7-10]。而这一点也是十分自然的，因为图模型（Graph）中的节点与边的结构能够表达非结构化的信息，这其中就包括了之前提到的用户行为信息，例如，在社交网络中用户之间的社交关系、电商平台中用户

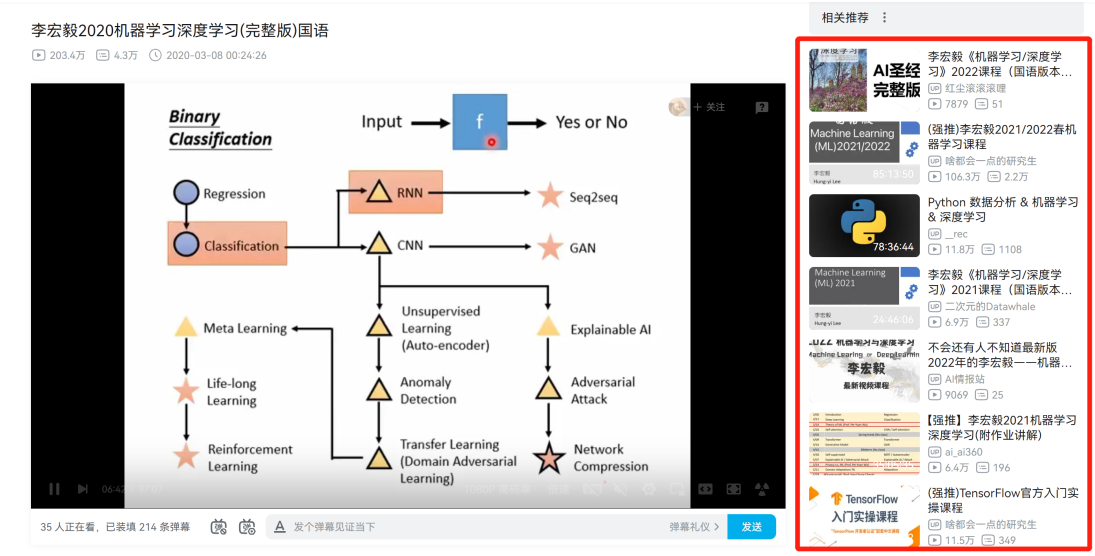

图 1-1 **视频网站 bilibili 的视频播放页面** 右侧红框部分为相关视频推荐列表。在机器学习的课程视频的相关推荐页面下，展示了此课程的其他版本以及相关编程语言教程等视频，这些可能是潜在的用户会点击的对象

对物品的浏览和购买行为记录等。图模型通过利用节点和边有效地表示实体之间的关系网络，将此作为数据的载体，能够更高效地在数据中挖掘到有价值的信息，并利用这些信息构建更有效的推荐系统。因此，图神经网络也越来越多地被用在推荐系统的各种场景。

图神经网络所依赖的图的建模往往有着非标准化的多样选择，图模型构图的多样性也给深度学习模型带了多样化的结构和适用场景。图根据图中边的方向可以被分为有向图和无向图，根据图节点性质的相同与否可以被分为同质图和异质图，根据图是否允许重边可被分为简单图和多重图，根据图中同一边可连接的节点数量则可分为一般图和超图等等。此外，图中的点或者边可以有属性或者无属性，边的属性中也可包含序列信息或时序信息等。针对推荐场景而设计的图模型也会有类似的多样的特点，例如，图节点可以表示用户或物品，抑或两者兼有，在此情况下，节点之间的连边则会因为连接节点类型的差异而产生出不同的含义，这样的图则有着“异构”的特征。从这一点出发，图模型的多样化建模既给图神经网络的结构带来了更多的可能，但同时由于目前没有通用的的构图标准，也给模型的设计者带来了挑战。如何有效率地通过图模型表达实体间的关系，很大程度也决定了图神经网络的效率。

在这里，本文引入超图（Hypergraph）的概念。超图是一类特殊的图，其节点与一般的图无异，但不同的是，一般的图的边仅连接图中的两个节点，而超图中的边（称为超边）可以连接大于两个数量的节点，从而表示多个节点之间的关系。 图 1-2 对比了一般图和超图模型结构上的区别。超图模型中超边能够将多个实体相联系，因此，十分适合在推荐系统这类复杂的场景应用。例如，在音乐

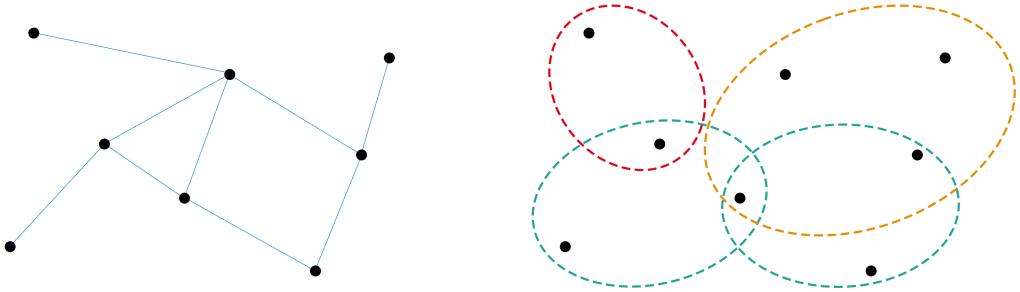

图 1-2 **一般图和超图** 左侧表示的一般图中，边（蓝色直线）连接两个节点。而右侧表示的超图中，超边（彩色的椭圆虚线）可以连接多个节点

平台上，图节点表示的实体可以有用户、歌曲、艺人、唱片厂牌等等。而这些实体间存在复杂的关系，如多首歌曲可以组成一张专辑、歌曲可以属于一个或多个艺人、用户可以收听歌曲和关注艺人等等。如果利用图神经网络对用户进行歌曲和艺人的推荐，很难用一般图将这些复杂的关系都表达出来。一种折衷的方式是用多个图分别表示这些关系再进行汇总，但这个方法不但仍然无法完整表达多个节点之间的相互联系，还可能出现不同的用户行为之间交互不够的问题，从而丢失重要信息。超图在此场景便有着强大的兼容能力，不同的超边不仅可以连接多个实体，而且能够存储超边本身的信息来表示这些实体之间具体的关系。同时，超图不仅仅只适用于此音乐平台场景，而是有着极强的兼容性。我们可以针对不同的场景任意定义不同的超边含义，使得图中能聚集所有需要的有价值的信息，使得需要被考虑的节点关系用同一张图完整地表达。然而，目前关于超图的节点表示学习的相关研究仍不完善，此外，基于超图的图神经网络鲜少被用在推荐系统相关的场景，这也是本文研究开展的一个重要背景。

在实际的推荐系统场景，用户和物品的规模往往达百万甚至千万乃至上亿。而一个可能会遇到的问题是，用户的某些行为数据有可能出现极其稀疏的情况。在这种情况下，以用户行为构建超边会使得整个超图也呈现稀疏的特征，而使得整个图的节点不能被充分连接，导致超图中的节点关系表达不足，超图节点表示无法得到有效训练。此外，构建超边由于基于图神经网络的对用户行为进行预测的模型的参数量往往与节点的数量至少呈线性关系，且需要多层的聚合和迭代的复杂运算，所以模型的有效训练要求与参数量规模相匹配的行为数据。因此，边稀疏也是在运用图神经网络时需要被考虑到的场景，否则模型有可能会出现泛化性不足的问题。

综上所述，基于超图的图神经网络在推荐系统上的应用是一个亟待研究的方向。本文将从这个出发点展开后续研究，提出通用场景下利用超图模型对用户和物品的关系建模的方法，以及基于此超图的神经网络点击率预测的模型。本文还给出了在没有充足用户行为数据时的模型优化方法。

1.2 研究难点

本研究的难点包括两方面。一方面是对于常规的推荐场景，如何利用超图对用户行为数据进行建模，基于超图的图神经网络如何聚集和更新节点表示，以及如何端到端地训练基于超图的图神经网络点击率预测模型。另一方面是对于某些特殊的场景，例如在用户行为数据稀疏的情况下如何提高模型的泛化性，以及模型如何在大规模数据上扩展。本节以下的部分将详细阐述四个这两方面的研究中的难点。

- **用户行为数据的超图建模** 如何对用户行为建模一直是推荐系统中的难点之一。图模型的特点即是能够表示节点之间的非结构性的拓扑关系，因此能够很好的被应用在多种场景，也被越来越多地用在推荐系统的建模中。但一般图的表达能力仍然有一定局限，而超图则进一步提升了图模型的表达能力，更加丰富了边能够表达的含义。然而，超图模型要如何最大限度保留有价值的交互信息，以及如何针对不同的业务场景对用户行为数据进行建模仍然是一个有挑战性的问题。
- **超图神经网络节点表示的聚合与迭代** 不同于一般图神经网络直接从邻居节点聚合和更新节点表示，超图中的“邻居”是一个定义不同的概念。处于同一超边的图节点可视为邻居，然而不同超边中的邻居对于图节点的意义显然是不同的。因此，原有的图神经网络所采用的方法无法直接被用于超图节点的表示学习中。一个比较简单的做法是不加区分地从邻居节点聚合信息，但这种做法有可能会丢失部分重要信息，比如跟超边直接相关的信息等。因此，如何对不同超边的邻居加以区分并聚合和迭代表示也是基于超图地神经网络需要考虑的问题之一。
- **用户行为数据的稀疏性** 在某些推荐场景下，通常有着用户行为数据稀疏的情况。例如对于用户的购买行为，往往用户基于大量浏览商品后才会有少量记录。在用户行为数据和物品的数量不相称的情况下，直接运用基于超图的图神经网络模型可能会造成一些问题。在上面这种情况下，以购买行为构建的超边的数量则会相较于用户和物品节点稀疏。如果仅以此方法构建超边，可能会使得图连通性极差。如何在此场景下仍然利用超图对用户行为信息进行建模也是需要被解决的问题。

1.3 论文的内容与意义

本文基于前面的背景和研究难点，开创性地提出了基于超图注意力机制的图神经网络点击率预测模型 **HGAT-GNN** （**H**yper**G**raph **AT**tention **G**raph **N**eural **N**etworks）。

正如名字中体现的，此模型利用了超图对与推荐系统有关实体间的关系建模，用于捕捉信息在这些实体之间的流动。其中图模型用节点表示用户和物品，

将用户的历史行为作为超边，并记录相应的动态信息。对于超图的节点隐层表示的学习，模型利用注意力机制从与节点相关的超边聚集隐层表示。以此构建的超图注意力层使用实体的静态特征初始化超图的节点表示，使节点能从相邻超边中获取有效的信息进行本身表示的迭代。之后利用节点的动态特征和注意力机制和从邻居节点更新隐层表示。而对于超边稀疏的情况，本文提出了将对应的稀疏节点类用动态特征表示的优化方法，使得模型在此场景具备更强的泛化性。将从超图中学习得到的节点表示直接作为特征输入至后续的预测网络中，可以为预测提供更多的特征信息。

从后续实验可以得出，此模型很好地利用了实体间的交互信息，并基于此提升了点击率模型的效率。实验部分将本文模型与多个基线模型进行了对比，不仅充分证明了模型的有效性，同时通过超参数分析和消融实验多方面验证了模型设计的合理性。最后，本文还详细总结了所做工作，并提出了今后工作的可能开展方向。

综上所述，本文做出的贡献有以下几点：

- 提出了基于超图注意力机制的图神经网络点击率预测模型 **HGAT-GNN**。此模型具有特征挖掘和交叉能力。与未考虑交互信息的深度神经网络模型和图神经网络模型相比在点击率预测任务表现较好，能有效在推荐场景对用户和物品的交互行为进行建模并进行点击率预测。
- 针对推荐场景的行为数据的稀疏问题提出了稀疏超图下的优化方法，即利用动态特征的相关计算替代节点的隐层迭代计算。不仅很好地解决了数据稀疏可能带来的模型过拟合问题，还能借此减少模型参数量，提高模型泛化性和训练效率。
- 一系列实验表明，基于超图注意力机制的图神经网络点击率预测模型 **HGAT-GNN** 在两个真实数据集上表现超过现有的相关基线模型。

1.4 章节结构

本文之后的章节安排如下：

第 2 章介绍了与本文相关的工作，包括图神经网络和推荐系统两部分，分别详细介绍了相关领域的经典与最新工作。

第 3 章介绍了本文提出的基于超图注意力机制的图神经网络点击率预测模型，详细地描述了每层结构的具体设计和计算方法，并给出了稀疏超图下的优化方法以及训练相关的事宜。

第 4 章介绍了实验数据集和模型在数据集上的实验结果，以及模型有关的超参数测试和分析以及消融实验。

第 5 章总结了本文并提出了之后工作的展望。

文章最后还包含了参考文献和致谢部分。

第 2 章 相关工作

本章介绍了与本文研究相关的工作：第一节介绍了图神经网络相关的研究；第二节先分别从传统和深度学习推荐模型介绍了两类模型的发展和继承，之后详细介绍了图神经网络点击率预测模型这一较新分支的最新研究成果。

2.1 图神经网络

图神经网络（Graph Neural Network, GNN）被提出用来解决图上节点的向量表示[11]，即图嵌入（Graph embedding），的问题，以提取图中的结构和交互信息。其中，图模型中的节点和边对实体和他们的关系进行建模。早期的工作中，DeepWalk[12]受到自然语言处理的启发在图上随机游走学习序列表示。LINE 算法[13]在此基础上加以改进，能够聚合图上一阶和二阶的结构信息。Node2Vec[14]也在此基础上做了改进，提出了有偏置参数的随机游走方法。HetGNN 模型[15]能够在异构图的学习节点嵌入。阿里巴巴提出的 EGES 模型给出了超大规模数据集上商品嵌入的解决方案[16]。

图神经网络的关键在于节点如何通过他们的邻居聚合信息并迭代他们自己的隐层表示。对于这个问题，一些经典的方法被相继提出。门控图神经网络（Gated Graph Neural Networks, GGNN）使用门控循环单元（Gated Recurrent Unit, GRU）[17]来进行迭代。图卷积网络（Graph Convolution Networks, GCN）[18]在图上对领域进行卷积提取局部特征。GraphSAGE 模型[19]加入了对邻居节点进行采样以及多跳邻居的聚合信息进行图嵌入的计算。图注意力网络（Graph Attention Networks, GAT）[20]利用注意力机制聚合邻域特征。Bai 等人提出并总结了 GCN 和 GAT 机制在超图中的应用[21]。

2.2 推荐系统

推荐系统对于互联网用户的增长有着强大的推动力。因此，近十多年来，学术界和工业界相关的研究发展迅速。推荐系统的主流模型按发展时间可以分为前深度学习时代和后深度学习时代[22]，即传统推荐模型和深度学习推荐模型。

2.2.1 传统的推荐模型

传统的推荐模型多基于相似度的计算，对特征信息的利用有限，但为之后深度学习模型提供了众多思路，奠定了基础。

主要的传统推荐模型包括协同过滤、逻辑回归、因式分解和集成模型四类。协同过滤算法[23]是最经典的推荐算法之一，由亚马逊公司提出，通过计算用户或物品之间的相似度进行同类推荐。矩阵分解算法[24]的提出解决了协同过滤算法的泛化性弱以及头部效应明显的问题。随着机器学习技术进入主流视野，逻

辑回归也被用于点击率的预测中。但上述模型都无法进行特征之间的交叉，FM 模型[25]和基于此提出的 FFM 模型[26]解决了这一问题。集成学习中 GBDT 模型则进一步提高了特征交叉的维数，而由 Facebook 提出的 GBDT+LR 的解决的方案[27]推动了推荐系统的特征工程自动化的进程。

2.2.2 深度学习推荐模型

深度学习推荐模型在 2015 年后已逐渐成为主流，有着表达能力和数据挖掘能力强、模型灵活性大、能根据不同的业务场景适应等众多优点。进入深度学习时代以后，深度学习模型凭借着更强的拟合和表达能力，能够灵活的针对不同场景进行调整契合，对推荐系统产生了变革性地影响。

在推荐系统的召回阶段，微软率先提出了 DSSM 双塔模型[1]用于用户和物品向量的分别学习和融合，Google 也相继提出了双塔 DNN 模型[3]，利用深度学习模型进行协同过滤的模型 NeuralCF 于 2017 年被提出[2]。微软于 2016 年发表的 Deep Crossing 模型[28]是一个深度学习应用于推荐系统排序模型的经典实践，Google 同年提出的 Wide&Deep 模型[4]也对业界产生深刻影响。在 FM 模型基础上提出的 FNN[29]、DeepFM[30]、NFM[31]等模型在深度模型的基础上提高了特征交叉的非线性表达能力。对于推荐模型中的多任务问题，ESSM[32]、DeepMCP[33]、MMoE[34]等模型也被相继提出。深度模型逐渐往复杂化、多样化的方向发展：阿里巴巴提出的 DIN[35]，将注意力机制引入深度学习模型，之后接着提出了 DIEN[5]，使得模型能够处理序列化的数据；上海交通大学提出的 UBR4CTR 模型[6]则解决了长序列建模的问题；将强化学习用于推荐的 DRN 模型[36]也被提出；而将卷积神经网络用于推荐的模型包括 FGCNN[7]、CSCNN[8]等。

2.2.3 图神经网络点击率预测模型

图深度学习模型近年来也被直接用于推荐系统中特征交叉和点击率的预测中，代表模型有以下。GCMC 模型[37]通过图的自编码器实现对图节点嵌入的学习。中科院提出的 Fi-GNN 模型[38]利用图神经网络进行特征的高阶交叉，并端到端地进行 CTR 预测。阿里巴巴提出的 PCF-GNN 模型[9]能够完成进行自监督预训练学习图嵌入的任务。华为提出的 DG-ENN 模型[10]在用户物品交互图上用交互关系捕获特征关系。微信提出的 GraphTR 模型[39]能够利用 FM、Transformer、GraphSAGE 等多种机制相结合进行视频标签的排序任务。

对于基于超图的图神经网络推荐模型，直到最近才有少量研究。如结合社交媒体与音乐内容的超图神经网络音乐推荐模型[40]，以及针对电商场景的多目标超图推荐模型[41]。然而，这些超图的建边逻辑都是基于人为定义的关系，而利用用户行为数据进行构图的端到端图神经网络点击率预测的研究尚缺。这也是本文研究的一个贡献之一。

第 3 章 方 法

本章详细介绍了基于超图注意力机制的图神经网络点击率预测模型的方法细节：第一节给出了此方法的问题定义和相关的表述总结；第二节逐层给出了模型的设计架构，包括超图注意力层中如何对超图节点的隐层表示进行初始化、如何利用注意力机制聚合超边信息以及迭代超图节点表示、如何在稀疏超图下如何对不同的节点类型和特征类型进行优化、以及后续的预测网络中的自注意力层和多层感知器。最后还给出了模型训练用到的相关设置。

3.1 问题定义

不失推荐场景点击率模型的一般性，假设数据集中每条数据包含 k 个特征 $\boldsymbol{a_1}, \boldsymbol{a_2}, ..., \boldsymbol{a_k}$，其中第 i 个特征 $\boldsymbol{a_i}$ 来自于用户特征集 $\mathbf{A_U}$，物品特征集 $\mathbf{A_I}$ 和上下文特征集 $\mathbf{A_C}$ 的其中之一，以及每条数据对应的标签 $y \in \{0, 1\}$，分别表示用户对物品的未点击和点击的行为。点击率预测（CTR prediction）的任务为利用历史信息，在给定的特征条件下预测 $\hat{y}$，即用户对于物品发生点击行为的概率。此定义适用于大部分推荐场景，在此定义下，本文提出的超图构建方法和图神经网络点击率预测模型也有较强的一般性。

基于上段描述的场景和数据集可构建超图 $\boldsymbol{\mathcal{G}} = \{\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}\}$。其中 $\boldsymbol{\mathcal{V}} = \{\boldsymbol{\mathcal{U}} \cup \boldsymbol{\mathcal{I}}\}$ 为点集，包含用户集 $\boldsymbol{\mathcal{U}}$ 和物品集 $\boldsymbol{\mathcal{I}}$，分别对应的特征集合为 $\mathbf{A_U}$ 和 $\mathbf{A_I}$。超边集 $\boldsymbol{\mathcal{E}} \subseteq \{\boldsymbol{\mathcal{V}}^2 \cup \boldsymbol{\mathcal{V}}^3 \cup ...\}$，对应特征集合 $\mathbf{A_E}$。

由于模型的需要，本文在这里将物品和用户特征集分为自始至终不会发生更改的静态特征集 $\boldsymbol{A_I^s}$（如名称、属性等）和动态特征集 $\boldsymbol{A_I^d}$（如浏览量等）。根据定义，有 $\boldsymbol{A_I} = \boldsymbol{A_I^s} \cup \boldsymbol{A_I^d}$。同理，也有 $\boldsymbol{A_U} = \boldsymbol{A_U^s} \cup \boldsymbol{A_U^d}$。

超图注意力层的任务即为利用给定的信息构建映射

$$Hypergraph\ attention\ layer : \boldsymbol{\mathcal{V}} \longrightarrow \boldsymbol{H},$$

并将 $\boldsymbol{H}$ 用于替换或补充点击率预测中对应的用户或物品特征，作为后续预测网络的输入，以提升预测的有效性。

3.2 模型结构

为方便描述，此节将问题进一步明确。但不失一般性，本文提出的方法也适用于其他场景。

考虑相关物品推荐的场景，用户 $u \in \boldsymbol{\mathcal{U}}$ 在源物品（source item）$i_s \in \boldsymbol{\mathcal{I}}$ 的详情页的相关物品栏浏览到了目标物品（target item）$i_t \in \boldsymbol{\mathcal{I}}$，模型需要预测用户点击目标物品的概率 $\hat{y}$。

上述提及的相关符号总结如 表 3-1 所示。

表 3-1 符号与描述总结

符号	描述
$\mathcal{G}$	超图
$\mathcal{V}$	超图点集
$\mathcal{E}$	超图边集
$\mathcal{U}, \mathcal{I}$	用户和物品集合
$\boldsymbol{A_U}, \boldsymbol{A_I}, \boldsymbol{A_C}, \boldsymbol{A_E}$	特征集合
$\boldsymbol{A_U^s}, \boldsymbol{A_I^s}$	静态特征集合
$\boldsymbol{A_U^d}, \boldsymbol{A_I^d}$	动态特征集合
$\boldsymbol{H}$	超图节点表示

3.2.1 超图注意力层

根据定义，超图中的节点即为用户与物品。我们将超图中的超边定义为一个子点集之间节点的相互关系，可以为两个或多个用户之间的交互（如关注、私信、群聊等），或多个物品的关系（如同属一个分类、在一定时间段内与同一用户交互等），或若干个用户与物品之间的交互（如用户和收藏过的物品、物品和收藏的所有用户等）。有关超边的特征存储于 $\boldsymbol{A_E}$，其中包括了超边对应的行为类型以及信息，如发生时间等。

为简单起见和适应实验部分用到的数据集，本文将超边集定义为点击行为集 $\boldsymbol{\mathcal{E}} = \{(u, i_s, i_t)_m | m = 1, 2, ...\}$，其中第 m 条边 $e_m = (u, i_s, i_t)_m$ 为一次用户点击行为，表示用户 u 在 i_s 物品的相关物品页面发生了点击 i_t 的行为，其对应的特征为 $\boldsymbol{A_{E_m}}$，即边 e_m 所对应的特征边集特征。

易得，在上述定义下有 $\boldsymbol{A_C} = \boldsymbol{A_E}$。

对于超图中的每个节点 $\boldsymbol{v_i} \in \boldsymbol{\mathcal{V}}$, 有对应的第 t 级节点隐层表示向量 $\boldsymbol{h_i^t} \in \mathbb{R}^D$, 其中 D 为表示向量的维度。所有节点的状态向量构成了超图的第 t 级点集表示

$$\boldsymbol{H^t} = [\boldsymbol{h_1^t}, \boldsymbol{h_2^t}, ..., \boldsymbol{h_{|V|}^t}], \tag{3.1}$$

而最后一层的节点隐层表示即为上一节要计算的节点通过超图注意力层映射的隐层表示 $\boldsymbol{H}$。

当 t 大于 1 时，第 t 级隐层表示由第 $t-1$ 级表示迭代而来，这部分内容在本节的后部分将详细阐述。当 t 为 1 时，有

$$\boldsymbol{h_i^1} = [\boldsymbol{h_i^0} \parallel \boldsymbol{W_{type(v_i)}^s a_i^s})], \tag{3.2}$$

其中 $\boldsymbol{h_i^0}$ 是节点 i 的初始化表示向量, $\boldsymbol{a_i^s}$ 是节点 i 的静态特征向量, $\boldsymbol{W_{type(v_i)}^s}$ 是对应的线性变化权重矩阵，$type(v_i)$ 指定了节点 i 的类型（用户或物品），$\parallel$ 表

示向量的连接。在初始化状态中，加入节点静态特征的目的是为了使得初始化向量包含节点的固有信息，而其动态特征我们将在之后的迭代步骤中与超边特征相融合。

接下来考虑节点隐层每级表示的迭代。由于超图中的边可能包含大于 2 个节点，所以如同一般图神经网络所采取的从节点的邻居聚合信息无法做到。一种可采取的变化是定义超图中广义上的“邻居”，即与此节点共同存在于同一超边的节点均为它的邻居，但这种做法有可能会丢弃超边的有效信息，使得不同超边中的点被同质化处理。为区分不同的超边，同时汇合与超边有关的信息，这里首先利用超边中的节点和本身特征对于每条超边 $e_m = (u, i_s, i_t)_m$ 计算超边状态

$$\boldsymbol{e}_{\boldsymbol{m}}^{\boldsymbol{t}} = \boldsymbol{W}_{\boldsymbol{type}(e_m)}^{\boldsymbol{e}}[\boldsymbol{h}_{\boldsymbol{u}}^{\boldsymbol{t}} \parallel \boldsymbol{h}_{\boldsymbol{i_s}}^{\boldsymbol{t}} \parallel \boldsymbol{h}_{\boldsymbol{i_t}}^{\boldsymbol{t}} \parallel \boldsymbol{W}_{\boldsymbol{type}(u)}^{\boldsymbol{d}} \boldsymbol{a}_{\boldsymbol{u}}^{\boldsymbol{d}} \parallel \boldsymbol{W}_{\boldsymbol{type}(i_s)}^{\boldsymbol{d}} \boldsymbol{a}_{\boldsymbol{i_s}}^{\boldsymbol{d}} \parallel \boldsymbol{W}_{\boldsymbol{type}(i_t)}^{\boldsymbol{d}} \boldsymbol{a}_{\boldsymbol{i_t}}^{\boldsymbol{d}} \parallel \boldsymbol{A}_{\boldsymbol{E_m}}], \tag{3.3}$$

其中，超边状态的计算融合了三部分信息：超边中节点上一次迭代的表示 $\boldsymbol{h}$（$\boldsymbol{h}_{\boldsymbol{u}}^{\boldsymbol{t}}, \boldsymbol{h}_{\boldsymbol{i_s}}^{\boldsymbol{t}}, \boldsymbol{h}_{\boldsymbol{i_t}}^{\boldsymbol{t}}$）、超边中节点在与此超边相关的动态特征 $\boldsymbol{a}^{\boldsymbol{s}}$（$\boldsymbol{a}_{\boldsymbol{u}}^{\boldsymbol{d}}, \boldsymbol{a}_{\boldsymbol{i_s}}^{\boldsymbol{d}}, \boldsymbol{a}_{\boldsymbol{i_t}}^{\boldsymbol{d}}$）、以及此超边的特征 $\boldsymbol{A}_{\boldsymbol{E_m}}$。权重矩阵 $\boldsymbol{W}_{\boldsymbol{type}(u)}^{\boldsymbol{d}}$、$\boldsymbol{W}_{\boldsymbol{type}(i_s)}^{\boldsymbol{d}}$ 和 $\boldsymbol{W}_{\boldsymbol{type}(i_t)}^{\boldsymbol{d}}$ 均只与节点的类型有关，对动态特征进行线性变化。最后，所有的向量连接起来并与线性变化矩阵 $\boldsymbol{W}_{\boldsymbol{type}(e_m)}^{\boldsymbol{e}}$（仅与超边类型有关）相乘，即得到最终的超边状态。

接下来，对于超图中每个节点分别聚合信息。与一般图神经网络不同的是，本文采取从与节点相邻的超边聚集状态，而不是直接由节点状态聚集。这里，本文使用注意力机制聚合超边状态，即

$$\boldsymbol{a}_{\boldsymbol{i}}^{\boldsymbol{t}} = \sum_{v_i \in e_m}^{m} Att(v_i, e_m) \cdot \boldsymbol{e}_{\boldsymbol{m}}^{\boldsymbol{t}}, \tag{3.4}$$

其中，$Att(v_i, e_m)$ 是超边 e_m 对于节点 v_i 的注意力分数，与超边状态相乘后累加得到节点 v_i 的临时状态 $\boldsymbol{a}_{\boldsymbol{i}}^{\boldsymbol{t}}$。注意力的计算公式如下所示

$$Att(v_i, e_m) = \frac{exp(\boldsymbol{W}_{\boldsymbol{type}(v_i)}^{\boldsymbol{a}} e_m^t)}{\sum_{v_i \in e_j} exp(\boldsymbol{W}_{\boldsymbol{type}(v_i)}^{\boldsymbol{a}} e_j^t)}, \tag{3.5}$$

对于超边 e_m，其对于节点 v_i 的注意力分数等于超边状态乘以仅与节点类型有关的线性变换参数 $\boldsymbol{W}_{\boldsymbol{type}(v_i)}^{\boldsymbol{a}}$，再通过 $softmax$ 函数得到。

在迭代的最后一步，用上一级状态 $\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{t}}$ 和临时状态 $\boldsymbol{a}_{\boldsymbol{i}}^{\boldsymbol{t}}$ 更新下一级状态 $\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{t+1}}$。

$$\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{t+1}} = Update(\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{t}}, \boldsymbol{a}_{\boldsymbol{i}}^{\boldsymbol{t}})$$

本文仍遵循简单起见的原则，将 $\boldsymbol{a}_{\boldsymbol{i}}^{\boldsymbol{t}}$ 通过随即丢弃层（dropout）后与上一级状态相加的到下一级状态，即

$$\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{t+1}} = \boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{t}} + dropout(\boldsymbol{a}_{\boldsymbol{i}}^{\boldsymbol{t}}). \tag{3.6}$$

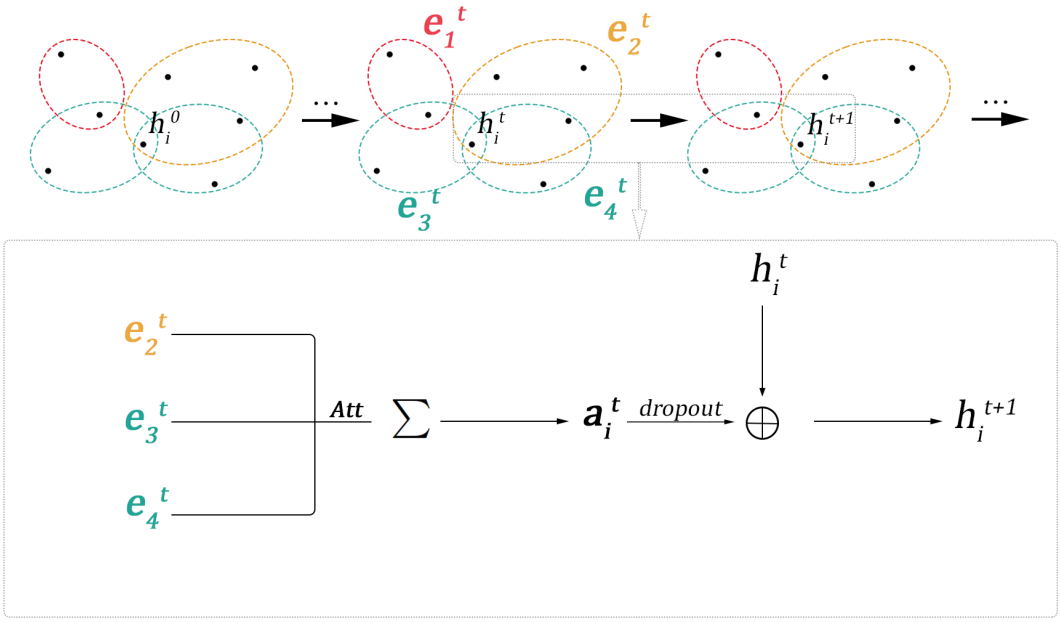

图 3-1 **超图注意力层示意图**

超图注意力层的架构如 图 3-1 所示。

3.2.2 稀疏超图下的优化方法

在推荐系统的实际应用场景，往往面临行为数据稀疏的问题。例如，如果在相关物品页面有几十甚至上百的目标物品，用户往往只会点击其中几个或甚至大部分时间没有点击行为。而大多数推荐场景往往有着用户数或物品数远远大于行为记录的特点。在此情况下，由 3.2.1 节方法构造的超图即存在节点数大于超边数的问题，整个超图呈现稀疏的特征。在此情况下，由于节点表示维数大于其对应的行为数据数量，容易使得节点表示的学习出现过拟合的问题。

为解决这个问题，本文提出了节点的动态特征表示方法。对于行为数据稀疏的节点类型（用户或物品，本文以物品为例），在每次节点表示的迭代计算中不再计算一一对应的向量，而仅通过动态特征乘以权重的形式出现在非行为数据稀疏的节点表示迭代计算中。即将公式 (3.3) 改为

$$\boldsymbol{e}_m^t = \boldsymbol{W}_{type(e_m)}^e [\boldsymbol{h}_u^t \parallel \boldsymbol{W}_{type(u)}^d \boldsymbol{a}_u^d \parallel \boldsymbol{W}_{type(i_s)}^d \boldsymbol{a}_{i_s}^d \parallel \boldsymbol{W}_{type(i_t)}^d \boldsymbol{a}_{i_t}^d \parallel \boldsymbol{A}_{E_m}], \quad (3.7)$$

公式 (3.7) 与公式 (3.3) 唯一的区别在于去除了 $\boldsymbol{h}_{i_s}^t$ 和 $\boldsymbol{h}_{i_t}^t$ 两项，即两个物品节点的隐层表示。这样虽然看似丢弃了部分信息，但是由于超边相对物品节点的数量稀疏，物品隐层表示中所包含的信息量不足以给提供更多的帮助，而两者

的作用可以被 $\boldsymbol{W}^{d}_{type(i_s)}$ 和 $\boldsymbol{W}^{d}_{type(i_t)}$ 两项替代。且对于不同超边，因为是动态特征，这两项提供的信息是不同的。

因为稀疏节点能够提供的动态特征已被相关节点考虑，在最后的预测网络层也不再采用这部分的节点表示。在这种情况下，对于物品隐层的计算也变得不必要，因此省去了图中大部分稀疏节点隐层表示的计算。后续的实验能够表明，这样的做法不仅能够大大减少模型的参数量和训练的计算量，提高训练效率，同时也能消除过拟合的影响，使得模型的泛化性更好。

3.2.3 预测网络

得到超图中的节点表示后，在原来的 k 个特征的基础上添加与本次展现有关的节点表示作为新的特征，或直接将原有的特征进行替换，进行后续的点击率预测。由于这不是本文研究的重点，这里只对所采用的模型作简单介绍。

自注意力层

首先，将 k 个特征向量和节点表示向量作为序列输入自注意力（Self attention）[42]层，进行特征之间的交互。具体地，我们有输入向量序列 $\boldsymbol{X} = \boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}$，其中 n 是序列长度，$\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}$ 均为长度为 d 的向量。

承接上文的定义，我们有

$$\boldsymbol{x_i} \in \boldsymbol{H} \cup \boldsymbol{A_U} \cup \boldsymbol{A_I} \cup \boldsymbol{A_C}, i \in [1, n], \tag{3.8}$$

之后用三个不同的线性变化权重矩阵 $\boldsymbol{W^Q}, \boldsymbol{W^K}, \boldsymbol{W^V}$ 分别对 $\boldsymbol{X}$ 进行变换，且 $\boldsymbol{W^Q}, \boldsymbol{W^K}, \boldsymbol{W^V}$ 具有相同的列数 d_k，分别得到变化后的矩阵 $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$，如下公式所示

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W^Q}, \tag{3.9}$$

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W^K}, \tag{3.10}$$

$$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W^V}, \tag{3.11}$$

序列特征向量间注意力分数的计算由此三个矩阵计算得到

$$\boldsymbol{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}, \tag{3.12}$$

其中，$\boldsymbol{K}^T$ 表示 $\boldsymbol{K}$ 的转置矩阵。将特征向量序列 $\boldsymbol{X}$ 通过注意力分数的变换得到交叉特征向量序列 $\boldsymbol{X'}$，即

$$\boldsymbol{X'} = \boldsymbol{X} \cdot \boldsymbol{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}), \tag{3.13}$$

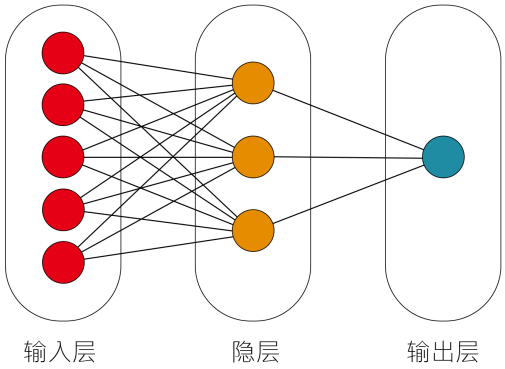

图 3-2 **多层感知器示意图**

这里，$\boldsymbol{X'}$ 是一个形状为 $n \times D$ 的矩阵，是由 $\boldsymbol{X}$ 通过自注意力层计算得来。在这一步使用自注意力层的目的在于能让从超图注意力层得到的隐层表示能充分与其他特征进行交叉，以得到更高阶的信息。这也是点击率预测网络的通用做法。从后一章的消融实验部分可以看出，加入自注意力层后模型的点击率预测明显优于直接将特征嵌入和隐层表示接入后续多层感知器的模型的效果。

将所有的交叉特征向量前后连接成长度为 $n \cdot D$ 的一维向量 $\boldsymbol{X''}$ 作为后续多层感知器的输入，即

$$\boldsymbol{X''} = [\boldsymbol{x'_1}||\boldsymbol{x'_2}||...||\boldsymbol{x'_n}], \tag{3.14}$$

多层感知器

将 $\boldsymbol{X''}$ 通过最后一层输出大小为 1 的多层感知器（MultiLayer Perception, MLP）[43]，即得到了最后的预测结果。图 3-2 给出了一个如所述的简单的多层感知器的结构。多层感知器将上层输出和下层输入相连接，对于第 t 层输入 $\boldsymbol{x_t}$，通过下面计算得到 $t+1$ 层输入 $\boldsymbol{x_{t+1}}$

$$\boldsymbol{x_{t+1}} = \sigma(\boldsymbol{W_t^{MLP}}\boldsymbol{x_t} + \boldsymbol{b_t}), \tag{3.15}$$

其中 $\boldsymbol{W_t^{MLP}}$ 和 $\boldsymbol{b_t}$ 是多层感知器第 t 层的权重参数。σ 是非线性激活函数，如 $sigmoid$ 和 $ReLU$ 等。此任务中，最后一层的激活函数须选择 $sigmoid$，以将结果的范围限制在 $[0,1]$ 内，同时也是假设标签服从多项式分布。

将 $\boldsymbol{X''}$ 通过多层感知器 MLP，得到最后的预测输出。

$$\hat{y} = MLP(\boldsymbol{X''}). \tag{3.16}$$

至此，整个模型的架构介绍完毕。基于超图注意力机制的图神经网络的架构图如图 3-3所示。

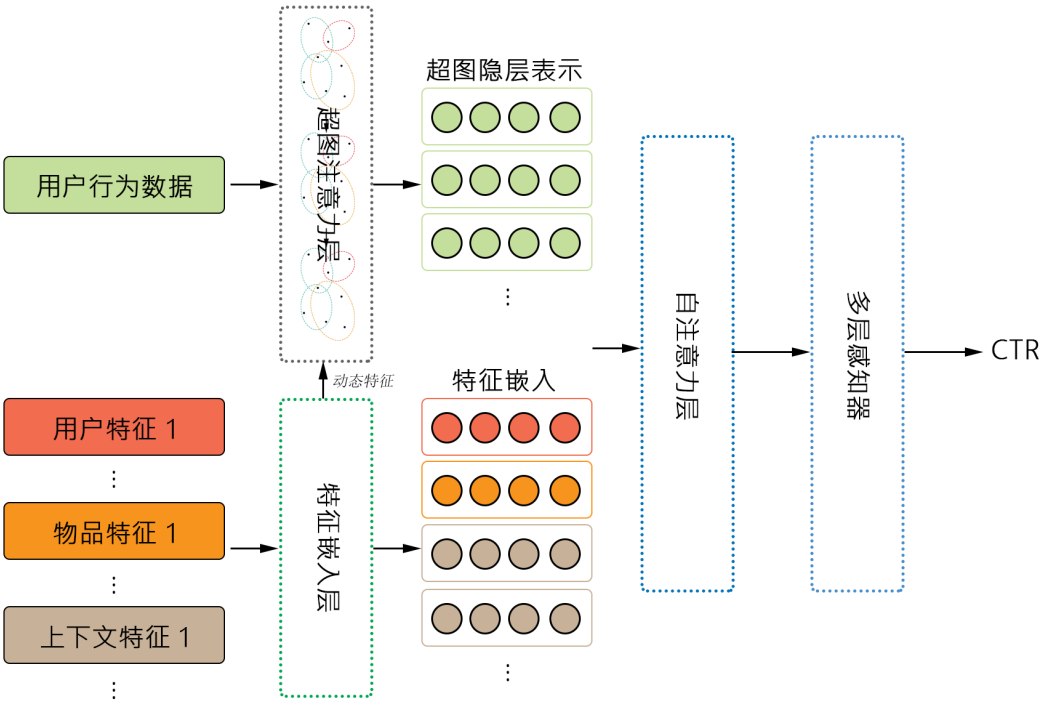

图 3-3 **基于超图注意力机制的图神经网络的架构图**

最后在这里总结基于超图注意力机制的图神经网络的架构，这是一个端到端的点击率预测模型。首先用用户的行为数据构建超图，作为超图注意力层的图模型。超图注意力层同时要接受特征通过特征嵌入层的嵌入（即隐层表示），并将其分为静态特征和动态特征表示分别作用于超图的隐层表示的初始化与迭代的步骤。超图注意力层最后输出超图节点的隐层表示，维度与用户特征、物品特征和上下文特征通过特征嵌入层的嵌入的维度一致。将这两部分表示和嵌入连接成序列之后，通过自注意力层进行交互，然后拼接得到一维向量。之后将此一维向量通过最后一层输出长度为 1 的多层感知器得到最后的点击率预测结果。

3.2.4 模型训练

此模型使用的损失函数为 Log 损失，定义如下：

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)). \tag{3.17}$$

其中，N 是样本的数量。当样本标签 y_i 为 1 时，括号中取前一项；当样本标签 y_i 为 0 时，括号中取后一项。

本文用 ADAM 方法[44]最小化损失函数。

此外，因为点击率预测任务数据集往往存在正负样本数量失衡的问题，且往往是负样本数量远大于正样本。因此，在训练之前需要对负样本进行随机采样，使其数量与正样本相当。

第 4 章 实 验

本节详细介绍了上章提出的模型在两个真实数据集上的实验：数据集一节介绍了实验开展的两个数据集的具体信息；实验设置一节给出了有关基于超图注意力机制的图神经网络点击率预测模型在两个数据集上部署的具体设置，以及相关的评估指标和对比基线模型的介绍；实验结果一节介绍了本文模型和基线模型在这两个数据集上的表现；超参数分析一节给出了对本文模型的超参数的选取的实验和分析，并给出了相应的选取建议；消融实验一节通过改变模型的一部分，以控制变量的方式验证模型各个结构设计的必要性和合理性。

4.1 数据集

本文采用视频网站 bilibili 推荐场景的真实数据集。数据集采集场景类似于 图 1-1 所示界面。用户在观看视频时或观看视频后，浏览相关推荐页面，有可能会对相关视频进行点击。每次点击或未点击行为均的相关信息会被记录，包括用户的信息、源视频的信息、目标视频的信息和其他上下文信息等。在实际运用中，大部分信息会被充分利用。但在此实验中，由于目的不是为了得到最好的在线结果，而是离线评估基于超图注意力机制的图神经网络点击率预测模型的有效性。因此仅采用少部分特征与基线模型进行对比，而不作为实际点击率预测的结果。当然，后续也可以将模型部署到完整数据上进行在线评估。

本文采取的数据集分为大小两个。因有涉密问题，部分数据进行模糊化处理。小数据集选取 2022 年 4 月 1 日至 202 年 4 月 7 日共计一周 7 天的用户数据，随机采样了约 3,000 个用户在某场景的共计约 600,000 条行为数据；大数据集选取 2022 年 3 月 1 日至 2022 年 3 月 31 日共计一月 31 天的用户数据，随机采样了约 8,000 个用户在某场景的共计约 16,000,000 条行为数据。其中，每条数据标签表示用户在源视频下浏览到目标视频的点击或未点击行为，特征字段包括用户 ID、源视频和目标视频的 ID、源视频和目标视频的分类特征各 4 项以及行为发生时上下文特征 2 项。

4.2 实验设置

本文采用第 3 章提出的基于超图注意力机制的图神经网络点击率预测模型在上述数据集上进行训练和预测。根据通用做法，将数据集按找 7:1:2 的比例划分为训练集、验证集和测试集。

对此数据集，建图的逻辑为为以用户和视频为节点，用户的点击行为为超边。鉴于观察到与视频有关的行为数据呈现稀疏的特征，故采用第 3.2.2 节提出的优化方法处理视频节点表示。

其他有关超参数和相关数据总结如 表 4-1 所示。

表 4-1 数据集和实验设置总结

	小数据集	**大数据集**
数据集及大小	约 600,000	约 16,000,000
覆盖时间	一个月（31 天）	一周（7 天）
用户数	约 3,000	约 8,000
视频数	30,000	约 800,000
视频分类特征类别数	(22, 11, 11, 11)	(22, 11, 11, 11)
上下文分类特征类别数	(8, 9)	(8, 9)
超边数量	约 30,000	约 800,000
特征嵌入层维度	32	32
超图注意力层注意力头数	1	3
超图注意力层隐层维度	8	16
自注意力层注意力头数	2	4
自注意力层隐层维度	16	16
多层感知器各层维数	(64, 32, 1)	(128, 64, 32, 1)
多层感知器各层激活函数	(relu, relu, sigmoid)	(relu, relu, relu, sigmoid)

4.2.1 评估指标

AUC (Area Under Curve) 是 ROC 曲线下与坐标轴围成的面积，衡量了模型正样本预测分数大于负样本预测分数的概率。数值越大代表预测模型表现越好。

AUC 相对提升率 衡量了模型 AUC 指标相对于基线模型的提升比例。数值越大代表预测模型表现越好。

Log 损失 衡量了模型预测概率与实际标签的距离。数值越小代表预测模型表现越好。

4.2.2 基线模型

LR（逻辑回归，Logistic Regression）进行了一阶特征交叉的线性模型。

FM[25]（因子分解机，Factorization Machine）进行了特征的二阶交叉，相较 **LR** 提升了特征交叉能力。

MLP（多层感知器，Multilayer Perception）利用深度神经网络（Deep Neural Networks, DNN）进行高阶的特征交叉。

CrossNet (Deep Crossing)[28] 将特征外积并连接而进行逐位的交叉。

Fi-GNN[38] 代表性的利用图神经网络进行特征交叉的点击率预测模型。但没有考虑交互信息。

表 4-2 实验结果

	小数据集			大数据集		
	AUC	AUC 相对提升率	Log 损失	AUC	AUC 相对提升率	Log 损失
LR	0.5000	13.64%	NaN	0.5000	22.28%	NaN
FM	0.5012	13.37 %	15.9045	0.5018	21.84%	4.4733
MLP	0.5426	4.72%	0.6960	0.5907	3.50%	0.6938
CrossNet	0.5539	2.58 %	0.6894	0.6007	1.78 %	0.6840
Fi-GNN	0.5501	2.39 %	0.6927	0.5997	1.95 %	0.6882
HGAT-GNN	**0.5682**	-	**0.6893**	**0.6114**	-	**0.6736**

4.3 实验结果

实验结果如表 4-2所示。可以看到，本文提出的模型在两个数据集的多项指标中取得了领先的结果。既证明了通过交互行为聚集信息的图深度网络在推荐任务上的有效性，又证明了超图模型具有强大的表达和特征交叉能力。

从图中的结果可以看出，**LR** 和 **FM** 两个模型在此数据集上的预测结果几乎接近随机（0.5）预测，不足以在此复杂的场景运用。而 **MLP** 和 **CrossNet** 两者引入了深层网络，能更好地进行特征交叉，因此取得了比前两者较好的结果。虽然 **Fi-GNN** 引入了图神经网络进行特征交叉，但可能由于其后续网络不如 **CrossNet**，其在此数据上的结果不如后者。虽然本文提出的 **HGAT-GNN** 模型没有致力于后续网络的结构设计，但由于其较好地利用超图上的图神经网络引入了用户行为信息，无论是在 AUC 还是 Log 损失上，得到了相较之前这些模型都更好的结果。

对于大小两个数据集，从 表 4-2 可以看出，以上所有模型在大数据集上的表现，无论是 AUC 还是 Log 损失，均好于小数据集。这除了是由于训练资料增多带来的提升之外，也是由于在大数据集上构建的超图要比小数据集上稠密。因此，大数据集上超图的节点表示包含的信息可能比小数据集上更丰富，也就给点击率预测提供了更多信息。而横向对比可以发现，**HGAT-GNN** 在两个数据集上相较于基线模型的提升各有千秋，但基本持平。因此可以说，无论超图的稠密程度如何，其给点击率预测提供的信息都对最后的点击率预测网络有所帮助。

4.4 超参数分析

实验中模型设置了众多超参数，超参数的选择通过控制变量的方法进行实验选择。本章以小数据集为例，通过分别调整 4 个超参数来获得分别的最好效果，并以此标准作为最后模型最后的参数选择。大数据集上也以同样的方法进行超参数的选择。有关不同超参数下模型的实验结果如 图 4-1 所示。

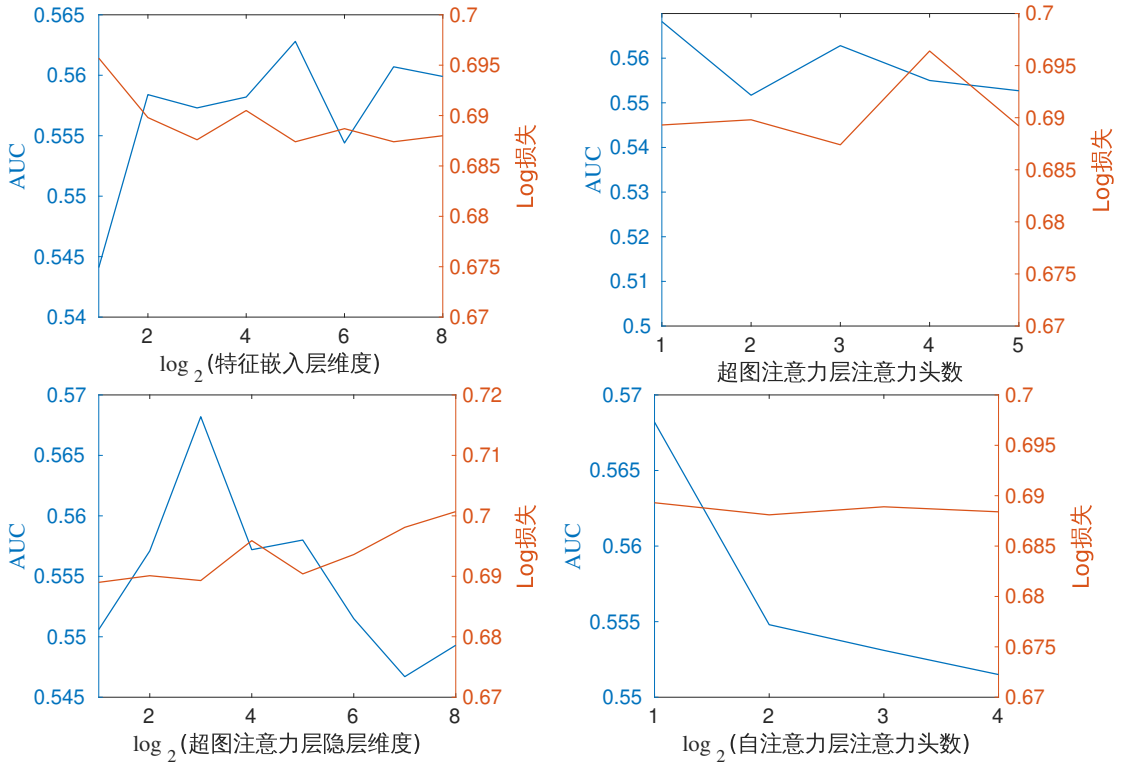

图 4-1 **超参数分析实验结果** 从左至右，从上至下分别为模型预测结果随特征嵌入维度、超图注意力层注意力头数、超图注意力层隐层维度和自注意力层注意力头数和输出维度的函数变化示意图

- **特征嵌入层维度** 特征嵌入层维度决定了特征嵌入能表示的信息量的大小。按照惯用做法，本文从小到大依次选取了维度为 2 次幂整数的维度大小进行模型预测。如 图 4-1 左上子图所示，随着嵌入层维度的增大，AUC 呈现整体上升的趋势，Log 损失呈现整体下降的趋势。但随着特征嵌入层维度达到一定值，AUC 和 Log 损失上升和下降的趋势逐渐放缓。而维度的幂次增加 1，训练所消耗的时间几乎也会翻倍。为了在训练效率和最终效果之间权衡，本文最终在小数据集上选取了 32 作为特征嵌入层的维度。
- **超图注意力层注意力头数** 超图注意力层注意力头数相当于超图注意力层利用注意力机制得到的不同的聚合信息的数量，不同的注意力头越多可能会得到越多不同的注意力分数，使得模型能够处理表示的异意性。但模型的参数量也与注意力头数呈现正比关系。从 图 4-1 右上子图结果中可以看到，此模型在小数据集中改变注意力头数对模型效果的影响较小，因此选择效率最高的 1 作为最后模型的超参数。但值得一提是，对于大数据集，当注意力头数取 3 时得到了明显较好的实验结果。
- **超图注意力层隐层维度** 超图注意力层隐层维度时超图的隐层表示的大小，维度越大隐层信息量越大。因其具体含义与特征嵌入层维度类似，这里不做过多赘述。根据 图 4-1 左下子图，AUC 曲线的形状类似于山峰，并在

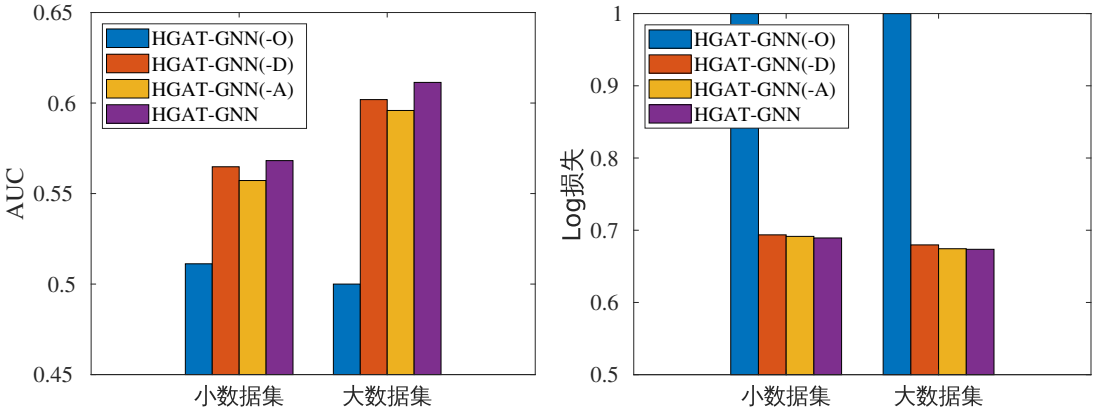

图 4-2 **消融实验结果** 两张图分别表示 **HGAT-GNN** 和其他消融模型在大小数据上的 AUC 和 Log 损失上的表现

横坐标为 8 时到达最大值，因此我们最终选取 8 为超图注意力层的隐层维度。

- **自注意力层注意力头数和输出维度** 因为最后要将不同注意力头的输出相连接作为后续特征输出，注意力头数和隐层维度的乘积须等于特征嵌入层维度，两者是一一对应的。 图 4-1 右下子图给出了模型预测结果随注意力头数的变化示意，本文选取表现最佳的参数组 (2×8)。

4.5 消融实验

本节根据模型的结构，针对性地去除部分设计，再用模型分别对大小两个数据上进行点击率预测。消融实验的目的是证明模型各个部分的设计是有效的。本文设计的多个消融模型的介绍如下。

- **稀疏超图下的优化方法（HGAT-GNN(-O)）** 去除了 3.2.2 提出的优化方法，仍然计算超图中物品节点的表示。可以从 图 4-2 可看出，在两个数据集上，模型均出现了不同程度的过拟合情况，其 AUC 接近 0.5，而 Log 损失则远远大于 1。可以说，在此场景下，这个优化方法是必需的。
- **动态特征（HGAT-GNN(-D)）** 在公式 (3.3) 中去除了动态特征项。如 图 4-2 所示，此消融模型的表现略次于原模型。可以因此得到结论，动态特征在超图隐层表示迭代中起到了重要作用。
- **超图注意力权重（HGAT-GNN(-A)）** 将用超图注意力层注意力机制改为平均加和。此消融模型的表现也次于原模型。这是由于注意力机制能使得有区分地进行表示的聚合和迭代。

通过上述三个消融模型的实验结果可以看出，**HGAT-GNN** 模型的各部分设计存在合理性，对模型最终的优良表现均产生了各自的贡献。

第 5 章 总结与展望

5.1 本文总结

本文提出了基于超图注意力机制的图神经网络点击率预测模型 **HGAT-GNN**。

针对于推荐系统复杂的业务场景，模型利用超图对用户的行为数据进行建模，能够有效地捕捉与推荐系统有关实体的交互关系，并依赖这种关系学习，利用注意力机制在超图上聚集和迭代超图节点表示，并以此作为实体特征与其他特征进行交叉后输入后续的网络模型，获得了相较于目前最新相关研究最好的点击率预测效果。

对于用户行为数据稀疏的场景，本文针对性地提出了稀疏超图下的优化方法，将对应超边稀疏的节点类型所对应的特征拆分为不会随行为变化的静态特征和随行为变化的动态特征，在节点表示的初始化中使用静态特征，而在迭代中不再计算这部分节点的节点迭代，而仅用其动态特征进行变换后用于其他节点的表示聚集和迭代。这种方法提高了模型的泛化性，同时减少了模型的复杂度，降低训练的计算量。

5.2 未来展望

在本文已有工作的基础上，之后的工作的方向可以集中于以下几点。

- **动态超图模型** 在实际场景中，超图模型所定义的图是不断在变化的。首先是不断会有新的用户和物品节点的加入，用户历史行为的积累也会导致超边集的不断变换。除此之外，用户特征，如用户的兴趣，以及物品的特征往往也是动态变化的。这些变化往往非常密集，使得图的结构在短时间内便会发生较大变化。在此情况下，为应对瞬息万变的业务实际，需要不断重新计算超图的节点表示，但这始终与变化的发生有着较大时间差。为尽量减少这种时间差，可以考虑动态地迭代超图节点表示，这部分研究尚残缺。
- **时序超图模型** 本文的模型未考虑一个重要的用户行为数据特征，即时序信息。但时序信息往往在推荐系统中起到了至关重要的作用，用户的行为特点往往是跟随时间相变化的。在超图上考虑超边的时序信息也是一个可能的提高节点表示有效性的方法。
- **在线实验** 本文模型在两个 bilibili 真实数据集上取得了较好的离线效果，但尚未投入线上使用。将模型部署线上还面临更大规模的计算量和存储量的问题，以及一系列工程问题。如果能成功部署线上并取得相应的效果提升，将更好证明模型的有效性。

参考文献

[1] HUANG P S, HE X, GAO J, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 2333-2338.

[2] HE X, LIAO L, ZHANG H, et al. Neural collaborative filtering[C]//Proceedings of the 26th international conference on world wide web. 2017: 173-182.

[3] YI X, YANG J, HONG L, et al. Sampling-bias-corrected neural modeling for large corpus item recommendations[C]//Proceedings of the 13th ACM Conference on Recommender Systems. 2019: 269-277.

[4] CHENG H T, KOC L, HARMSEN J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. 2016: 7-10.

[5] ZHOU G, MOU N, FAN Y, et al. Deep interest evolution network for click-through rate prediction[C]//Proceedings of the AAAI conference on artificial intelligence: volume 33. 2019: 5941-5948.

[6] QIN J, ZHANG W, WU X, et al. User behavior retrieval for click-through rate prediction[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 2347-2356.

[7] LIU B, TANG R, CHEN Y, et al. Feature generation by convolutional neural network for click-through rate prediction[C]//The World Wide Web Conference. 2019: 1119-1129.

[8] LIU H, LU J, YANG H, et al. Category-specific cnn for visual-aware ctr prediction at jd. com[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 2686-2696.

[9] LI F, YAN B, LONG Q, et al. Explicit semantic cross feature learning via pre-trained graph neural networks for ctr prediction[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 2161-2165.

[10] GUO W, SU R, TAN R, et al. Dual graph enhanced embedding neural network for ctr prediction[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 496-504.

[11] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.

[12] PEROZZI B, AL-RFOU R, SKIENA S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 701-710.

[13] TANG J, QU M, WANG M, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th international conference on world wide web. 2015: 1067-1077.

[14] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks [C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 855-864.

[15] ZHANG C, SONG D, HUANG C, et al. Heterogeneous graph neural network[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 793-803.

[16] WANG J, HUANG P, ZHAO H, et al. Billion-scale commodity embedding for e-commerce recommendation in alibaba[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 839-848.

[17] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[18] YING R, HE R, CHEN K, et al. Graph convolutional neural networks for web-scale recommender systems[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 974-983.

[19] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[J]. Advances in neural information processing systems, 2017, 30.

[20] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.

[21] BAI S, ZHANG F, TORR P H. Hypergraph convolution and hypergraph attention [J]. Pattern Recognition, 2021, 110: 107637.

[22] 王喆. 深度学习推荐系统[M]. 北京: 电子工业出版社, 2020.3: 010-011.

[23] LINDEN G, SMITH B, YORK J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet computing, 2003, 7(1): 76-80.

[24] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[25] CLINE A K, DHILLON I S. Computation of the singular value decomposition [J]. 2006.

[26] JUAN Y, ZHUANG Y, CHIN W S, et al. Field-aware factorization machines for ctr prediction[C]//Proceedings of the 10th ACM conference on recommender systems. 2016: 43-50.

[27] HE X, PAN J, JIN O, et al. Practical lessons from predicting clicks on ads at facebook[C]//Proceedings of the eighth international workshop on data mining for online advertising. 2014: 1-9.

[28] SHAN Y, HOENS T R, JIAO J, et al. Deep crossing: Web-scale modeling without manually crafted combinatorial features[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 255-262.

[29] ZHANG W, DU T, WANG J. Deep learning over multi-field categorical data[C]// European conference on information retrieval. Springer, 2016: 45-57.

[30] GUO H, TANG R, YE Y, et al. Deepfm: a factorization-machine based neural network for ctr prediction[J]. arXiv preprint arXiv:1703.04247, 2017.

[31] HE X, CHUA T S. Neural factorization machines for sparse predictive analytics [C]//Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval. 2017: 355-364.

[32] MA X, ZHAO L, HUANG G, et al. Entire space multi-task model: An effective approach for estimating post-click conversion rate[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 1137-1140.

[33] OUYANG W, ZHANG X, REN S, et al. Representation learning-assisted click-through rate prediction[J]. arXiv preprint arXiv:1906.04365, 2019.

[34] MA J, ZHAO Z, YI X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts[C]//Proceedings of the 24th ACM SIGKDD

International Conference on Knowledge Discovery & Data Mining. 2018: 1930-1939.

[35] ZHOU G, ZHU X, SONG C, et al. Deep interest network for click-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018: 1059-1068.

[36] ZHENG G, ZHANG F, ZHENG Z, et al. Drn: A deep reinforcement learning framework for news recommendation[C]//Proceedings of the 2018 World Wide Web Conference. 2018: 167-176.

[37] BERG R V D, KIPF T N, WELLING M. Graph convolutional matrix completion [J]. arXiv preprint arXiv:1706.02263, 2017.

[38] LI Z, CUI Z, WU S, et al. Fi-gnn: Modeling feature interactions via graph neural networks for ctr prediction[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 539-548.

[39] LIU Q, XIE R, CHEN L, et al. Graph neural network for tag ranking in tag-enhanced video recommendation[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 2613-2620.

[40] BU J, TAN S, CHEN C, et al. Music recommendation by unified hypergraph: combining social media information and music content[C]//Proceedings of the 18th ACM international conference on Multimedia. 2010: 391-400.

[41] MAO M, LU J, HAN J, et al. Multiobjective e-commerce recommendations based on hypergraph ranking[J]. Information Sciences, 2019, 471: 269-287.

[42] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[43] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning internal representations by error propagation[R]. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[44] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.