硕士学位论文

基于异质信息网络的跨领域推荐系统

Heterogeneous Information Network Embedding based Cross-Domain Recommendation System

作者：尹姜谊

导师：郭宇春

北京交通大学

2020 年 5 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：尹善谊　　导师签名：

签字日期：2020 年 6 月 1 日　　签字日期：2020 年 6 月 1 日

学校代码：10004 密级：公开

北京交通大学

硕士学位论文

基于异质信息网络的跨领域推荐系统

Heterogeneous Information Network Embedding based Cross-Domain Recommendation System

作者姓名：尹姜谊 学　　号：17120156

导师姓名：郭宇春 职　　称：教授

学位类别：工　学 学位级别：硕士

学科专业：通信与信息系统 研究方向：信息网络

北京交通大学

2020 年 5 月

致谢

本论文的研究工作是在我的导师郭宇春教授的悉心指导下完成的。郭宇春教授严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。从课题的选择到项目的最终完成，郭宇春老师都始终给予我细心的指导和不懈的支持。郭老师不仅在学业上给我以精心指导，同时还在思想、生活上给我以无微不至的关怀，在此谨向郭老师致以诚挚的谢意和崇高的敬意。

感谢实验室所有老师。陈一帅老师，赵永祥老师、李纯喜老师、郑宏云老师和张立军老师的帮助和关怀使我能顺利完成研究生阶段的学习。特别是要对陈老师一直以来对我的鼓励和支持，表示诚挚的谢意。

另外，在实验室的学习和工作中，李俊峰师兄、陈滨师兄、张大富师兄、唐伟康师兄、盛烨师姐等师兄师姐们帮助我解决了很多疑惑，于兹灏、冯梦菲、戚余航等同学也给予了我很多帮助，在此向他们一并表达我的感谢之意。

最后，特别感谢一直给予我无尽的付出和支持的家人，正是来自他们多年的默默奉献，才使得我顺利的完成学业，成为社会的有用之才。

摘要

在互联网技术飞速发展的今天，个性化推荐系统已经在人类生活中扮演越来越重要的角色。传统推荐系统的个性化实现需要大量的用户行为信息。当用户和项目的互动行为数据稀疏甚至缺失时，个性化推荐系统的效果会受到不利影响，产生冷启动问题。

目前，为解决用户行为数据稀疏造成的推荐系统冷启动问题，主要的研究方向有基于异质信息网络的推荐和跨领域推荐。前者通过引入异质网络信息来补充用户项目互动行为数据的不足。可是，由于目标域数据本身不够丰富，并且基于网络结构的用户和项目的表达不能同时很好地提取个性化特征，异质信息带来的提升十分有限；后者引入数据相对密集的辅助信息，通过对辅助域知识的迁移和整合来提高模型的效果，但因为辅助知识的类型相对单一，模型的适应性和可扩展性较弱。

针对上述问题，本文提出结合基于异质信息网络和跨领域的推荐算法，同时引入异质信息和跨领域信息来解决用户行为稀疏导致的冷启动问题，并分别针对评分预测和 Top-K 列表推荐两种个性化推荐场景，提出了基于异质信息的跨领域推荐算法，最后通过实验基于真实数据对算法进行了有效性评估。具体贡献如下。

(1) 针对个性化推荐中由于行为数据稀疏造成的冷启动问题，本文提出采用基于异质信息网络的跨领域推荐算法。该算法采用标签作为桥梁，构造跨领域异质信息网络，同时引入异质信息和跨领域信息作为辅助信息，将基于异质信息网络的推荐和跨领域推荐进行有机结合，实现推荐算法的融合。

(2) 针对评分预测场景，提出基于网络表征的融合算法框架 HecRec 以挖掘复杂多样的跨领域异质信息。同时，为了避免了信息间冲突造成知识的负迁移情况，采用了“立交桥式”向量处理方法，以最大化融合框架的优势，提高推荐效果。实验证明，融合算法框架的绝对平均误差为 0.6384，比相关工作中表现最优的算法降低了 2.7%。

(3) 针对 Top-K 列表推荐场景，提出基于嵌入传播层的融合算法模型 EPCDRec 进行跨领域异质信息的挖掘，并在端到端模型中实现推荐效果的提升。融合模型在真实数据集上的准确率、召回率和归一化折损累计增益分别达到了 0.13、0.16 和 0.22，比相关工作中表现最优的算法分别提升了 2.8%、1.2%和 3.0%。

图 14 幅，表 6 个，参考文献 52 篇。

关键词：个性化推荐；冷启动；异质信息网络；跨领域推荐

ABSTRACT

Along with the high-speed development of information technology, personalized recommendation system begins to play an extremely indispensable role in our daily life. Traditional recommendation systems require a large amount of interaction information between users and items to make the system more personalized. As a consequence, if the interaction information is inadequate, the performance of the system will be affected, which is known as the cold-start problem in personalized recommendation systems.

Now days, there are two main research directions to solve the problem, heterogeneous information network(HIN) embedding based recommendation and cross-domain recommendation. The former one improves by introducing heterogeneous information to supplement the lack of interaction information. However, due to the insufficient data in the target domain, the improvement is limited. And also, the expression of users and projects based on the network structure cannot simultaneously extract personalized features. The latter one, cross-domain recommendation introduces auxiliary information from another domain, which has relatively dense dataset, and improve the effect of the system by transferring and integrating the auxiliary knowledge. But the type of the transferred knowledge is always single, so that the adaptability and scalability of the system are not good.

In response to the above problems, we propose to combine heterogeneous information networks and cross-domain recommendation framework and to introduce heterogeneous information and cross-domain information to solve. Aiming at the two common recommend scenarios, score prediction and Top-K list recommendation, we propose HIN embedding based cross-domain frameworks respectively, and evaluate the effectiveness of them on real datasets. The specific contributions are as follows.

(1) To solve the cold-start problem, this paper proposes to use heterogeneous and cross-domain information. We use tags as bridge to construct cross-domain heterogeneous information network, combine heterogeneous information network and cross-domain recommendations organically to achieve a better performance.

(2) In the score prediction scenario, a fusion framework named HecRec is proposed to mine complex and diverse cross-domain heterogeneous information. Moreover, in order to avoid the knowledge conflicts conveyed by different meta-paths, we adopt the conception of "overpass" to process original embeddings. The experiments we conduct show that the absolute average error of HecRec is 0.6384, which is 2.7% reduced than the

related work.

(3) In the Top-K list recommendation scenario, we propose the fusion algorithm model EPCDRec to mine cross-domain heterogeneous information, which is based on embedding propagation layers and capable to study node embeddings characterizing both network structure features and personalized features in an end-to-end model. The accuracy rate, recall rate and cumulative normalized gain of the fusion model attain 0.13, 0.16 and 0.22 respectively, improved by 2.8%, 1.2% and 3.0% compared with related work.

14 figures, 6 tables, and 52 reference articles are contained in the dissertation.

KEYWORDS：Personalized recommendation; cold-start; heterogeneous information network; cross-domain recommendation

目录

1 引言

1.1 研究背景和意义

互联网科技的发展使人类获取的信息量大增。面对网络环境中持续增加的海量信息，如何进行数据处理和筛选，在尽量短的时间内获得目标信息，成为亟待解决的问题。为应对人们的日常需求，各种互联网信息服务层出不穷，社交媒体、电子商务、广告等线上的个性化推荐系统（Personalized Recommendation System）[1]在日常生活中加剧渗透，帮助进行目标信息的高效获取，不仅改变了人的生活方式和行为习惯，也加速了社会财富的积累。个性化推荐系统通过综合相关用户和项目的属性，以及用户和项目之间的交互行为等信息，对用户的个性化偏好进行分析和预测。常见的推荐系统包括主题交流平台（电影、图书、短视频等）、社交平台、购物网站、招聘平台等。个性化推荐系统的普遍使用涉及到生活的各个方面，用户对系统效能的要求也日益提高。

个性化推荐系统中存在两大主要场景：评分预测和 Top-K 列表推荐。评分预测即为用户对任意项目的实际评分进行预测。在这样的应用场景中，系统的目标是尽可能的使预测评分接近真实的评分值。模型需要用户对项目的显式反馈信息，即具体的喜欢或者讨厌程度的量化情况。模型的目标是对整体评分进行预测，不仅是对用户喜欢的项目，也包括用户不喜欢和讨厌的项目。这样的模型在项目维度也可以对单个项目进行整体评价。Top-K 列表推荐是针对用户生成长度为 K 的个性化推荐列表。这种场景下，系统往往只能获取用户的隐式反馈信息，比如购买、点赞或者浏览情况。模型的目标是将用户可能喜欢的项目挑选出来，实质上将用户的未接触项目按照是否喜欢分成了两类。和评分预测模型不同，它更侧重于将目标和非目标项目列表区分开。尤其是相对于 K 而言，项目集合本身规模较大，而 Top-K 推荐场景下只关注极少量的 K 个目标项目是否排序靠前，而不过分关心剩余的绝大多数项目之间的排名情况。评分预测场景主要出现于各类评分网站（如豆瓣），Top-K 推荐场景则较多出现在购物等网站中。很多情况下，同一网站中也会同时出现两种个性化的功能需求。图 1-1 给出了两种不同的推荐场景应用实例。

传统个性化推荐系统的实现，不论评分预测还是列表推荐，都需要大量的用户行为数据。一般来讲，数据量越大，模型效果越好。以推荐领域中经典的矩阵分解（Matrix Factorization）[2]算法为例，该算法对评分矩阵进行分解，分别获得用户和项目的嵌入式向量化表达，然后基于用户和项目间内积进行评分预测。这样的方法

在评分数据缺失的情况下，矩阵分解获得的向量化表示无法准确的刻画用户偏好和项目特征，预测效果会受到较大影响。而在推荐平台的投入使用初期，以及日常使用过程中新用户和项目的涌入，都会造成所需信息量相对稀疏，从而导致个性化目标难以实现的问题。这种问题也就是推荐系统中的冷启动[3]问题。如何有效的解决推荐中的冷启动问题，是推荐领域长期存在且较为困难的任务，对于提高个性化推荐系统的效果有重要意义。

图 1-1 两种推荐场景实例

Figure 1-1 Examples of two kinds of scenarios.

解决个性化推荐系统中的冷启动问题，也就是要解决用户项目之间交互信息不足的问题。常见的方案主要有两种：1）不同于传统推荐算法——更多的考虑系统中的用户和项目两类实体以及之间的关系，尽可能的挖掘除此之外系统中的其他相关的有效信息。比如电影推荐系统中，除用户和电影之外，影片类型、导演、演员等都可以作为有效信息对用户偏好和电影特征进行分析。而在考虑系统中多种不同实体及实体之间的复杂关系时，异质信息网络[4,5,6]，也就是包括多种节点和边的类型的网络结构，作为一种有效工具可以将异质系统表达为复杂网络，在网络拓扑层面对系统中包含的有效信息进行整合和利用，即基于异质信息网络的推荐。2）可以从其他数据相对密集的系统中提取有效的共享信息到目标推荐领域中来，通过知识迁移解决目标领域的数据稀疏问题，也就是跨领域推荐算法。

然而，现有的这两种方案都具有明显的不足：1）基于 HIN 的推荐技术一方面仍然受限于目标域数据的丰富性和稠密度，另一方面，网络结构特征不同于用户的个性化偏好特征[7]，因此如何将网络结构表征学习和推荐目标进行统一也是一个难点；2）跨领域推荐对辅助域信息挖掘不充分，迁移的知识种类较为单一，模型扩展能力相对较弱。因此需要设计新的方法和框架来弥补这两种方案存在的问题，改

善推荐在冷启动情况下的性能。

本文致力于研究基于异质信息网络的跨领域推荐算法，以解决数据稀疏问题从而提高个性化推荐效果。基于异质信息网络的推荐算法可以尽可能挖掘领域内复杂多样的信息，打破互动行为信息量对个性化推荐模型效果的限制；而从其他领域，即辅助域，迁移有效的共享信息到目标推荐领域中来，扩大了可用数据源的范围。本文针对两种常见推荐场景，即个性化评分预测和 Top-K 列表推荐，分别实现了融合跨领域异质信息的推荐框架，对解决数据稀疏造成的冷启动问题，提高推荐效果具有重要价值。

1.2 国内外研究现状

目前有大量研究工作尝试解决个性化推荐中数据稀疏的问题。下面将首先对传统推荐技术及其相关问题进行介绍，然后介绍不同解决方案的介绍来了解相关研究现状。

1.2.1 传统推荐技术及其问题

推荐系统通过从数据中挖掘有效信息来对用户个性化特征进行分析和预测。常见个性化推荐算法的核心思想是基于过滤的推荐。两种被广泛使用的过滤推荐策略分别是基于内容的过滤（Content Based Filtering）[1]和协同过滤（Collaborative Filtering）[1]。基于内容的过滤算法的核心假设为，如果一个用户偏好某一个项目，那么用户可能会喜欢其他相似的项目。基于内存和基于模型是协同过滤算法较为常见的两种类别。其中，基于内存的协同过滤可以分为为从用户角度和项目角度两种。而上述所有基于过滤的个性化推荐的相关算法都需要先验的互动信息来对个性化特征进行训练。由于更好地利用了的互动信息，在推荐中实现了偏好个性化，因此相对基于内容的算法[1,3,8,9]，协同过滤往往能获得更好的模型效果。可是协同过滤对于互动信息的依赖也导致了冷启动问题[10,11,12]的出现。

现有很多工作采用增加额外辅助信息的方式弥补个性化推荐算法中互动行为数据的不足。为了解决冷启动问题同时提高模型效果，Ma 等人[13]基于矩阵分解的算法[11,14]的思想，用社会网络信息引入两个社会关系规范项来约束矩阵分解的目标函数，将社会关系特征融入 MF 框架中；Nasiri[15]等人引入评分时间序列作为额外信息，将时间作为独立的维度，提出基于张量分解的推荐新方法；Rafailidis[16]等人提出了一种基于社交标签的个性化商品推荐新方法，通过相关性反馈机制来利用项目的内容信息和用户分配的标签信息。

1.2.2 基于异质信息网络的推荐技术

网络模型作为数据挖掘重要的建模方法和研究方向，被广泛应用于个性化推荐。推荐系统中的用户和项目作为实体节点，用户和项目之间的关系作为连边，构成网络结构。但传统的网络分析方法不适用于大规模网络情况，因此一些研究工作提出网络学习表征算法，将大规模的网络节点映射到低维的特征空间，即获得节点的嵌入式表达。嵌入式表达在特征提取中的优势使其能很好的应用于各类数据挖掘任务，推荐就是其中的一种。Deepwalk[17]结合了 random walk 和 skip-gram 学习网络表达；Grover[10]等人在有偏置的随机游走的基础上，提出了一种更灵活的表征学习框架；LINE[18]和 SDNE[19]能够提取二阶链路的相似度和邻接关系，还有一些工作[20, 21, 22]利用节点内容相关的信息来提高表征向量的鲁棒性。然而这些网络表达的方法多用于同质网络，即节点和边的种类不超过两种的网络结构。

近年来，作为个性化推荐研究领域最新出现的研究方向，异质信息网络（Heterogeneous Information Network，HIN），即节点或连边的类型超过两种的复杂网络结构，可以自然地将领域内的各种异质信息，进行整合建模。通过网络拓扑的角度分析网络结构中复杂的节点类型以及节点间的相互关系，探索用户互动行为信息及其他各种复杂的辅助信息，来挖掘用户和项目的特征并进行个性化预测和推荐。引入辅助信息，能改善互动行为信息量不足对个性化推荐效果的影响，即冷启动现象，提高整体模型效果。但是传统的同质网络表达算法[17, 18, 19]并不适用于异质信息网络，不同种类的节点在没有进行区别的情况下，会损失大量有效的异质信息，学习到的表征向量也会引入噪声。一些研究工作[4, 10, 18, 19, 51]提出，采用基于元路径的相似性度量方法来评价异质信息网络中节点的相似度，可以较好的综合系统中多样的辅助信息进行个性化特征的提取。Feng[23]等人提出了 OptRank 方法，通过利用社会标签系统中包含的异质信息来解决冷启动问题，并将元路径的概念引入到混合推荐系统中来适应异质信息网络的推荐场景；Yu[14]等人将基于元路径的相似度作为矩阵分解的正则项；Yu[15]等人利用异质信息网络中分析多类型节点的优势，提出了适用于隐式反馈的个性化推荐框架；Luo[16]等人提出了基于社交中的异质信息的协同过滤推荐方法；Shi[17]等人提出了带权重的异质信息网络，并设计了基于元路径的协同过滤模型，可以灵活的整合在个性化推荐中有效的异质信息；在 Zheng 等人的工作中[28, 30, 31]，用户之间和项目之间的相似度都基于不同语义的路径进行评估，并提出了基于双重正则化的矩阵分解来进行评分预测。

上述基于 HIN 的推荐方法都适应异质信息网络的数据特点，依赖基于元路径的节点相似性分析或嵌入式表达。除了用户和项目间，如评分等互动信息之外，这类算法可以引入领域内其他辅助信息，一定程度上改善对互动行为的依赖性，解决

推荐冷启动问题。但现有算法仍存在一些亟待解决的问题。一方面，采用基于异质信息网络的推荐算法，引入辅助异质信息，虽然能一定程度上改善用户行为数据的不足，但算法本质上仍受限于领域本身的数据量；另一方面，有工作[7]已表明，网络拓扑特征并不能直接表示用户的个性化偏好特征，无法直接应用与推荐的相关目标。所以目前基于 HIN 的推荐算法大部分是分模块多阶段实现的，不同模块分别承担网络结构的表征学习及用户偏好分析的功能，而难以实现端到端的推荐系统（End-to-end model）。

1.2.3 跨领域推荐技术

跨领域推荐（Cross-Domain Recommendation）通过引入其他辅助领域，即源域（Source Domain）的数据，来提升目标推荐领域（Target Domain）的模型效果。由于可用数据域范围的扩大，跨领域推荐技术在解决个性化推荐冷启动问题上也是一个重要的研究方向。

现有的跨领域推荐技术大多数从源域中学习或提取出目标域可用的有效信息，然后将提取出的信息补充应用到目标域中。在 CBT[32]及其优化算法[12, 14]中认为用户具有固有的评分模式，提出将从辅助域中学习的用户评分模式迁移到目标推荐领域进行评分预测的方法解决评分数据稀疏的问题；Chen[21]等人提出算法 TLRec，将不同领域重叠的用户和项目作为桥梁进行知识迁移；Li[14]等人提出迁移群组级别的评分模式到目标域中参与建模；Shi 等人[33]认为具有相同偏好的用户在标签的使用上同样具有相似性，提出 TagCDCF 框架，采用公共的标签内容作为领域间共享知识的桥梁进行知识迁移；Hao 等人[29]等人提出了 TagCDCF 的优化算法，该算法不仅依赖公共部分的标签，同时还对大比例的单领域标签的丰富信息进行挖掘。跨领域推荐作为挖掘辅助域信息的有效方法，一定程度上能够解决目标域数据稀疏产生的问题。但大多数跨领域推荐算法只进行某种特定种类的共享知识或公共关系，对辅助域有效信息的挖掘不够充分，模型的扩展能力也相对较弱。

1.3 研究内容及主要贡献

从上述研究现状分析中我们可以看出，针对冷启动问题的两类研究方案——基于异质信息网络的推荐技术和跨领域推荐技术，分别从域内和域外两个方面，引入不同来源的有效辅助信息。但两种研究方案各自存在问题也较为明显。基于 HIN 的推荐技术一方面仍然受限于目标域数据的丰富性和稠密度，另一方面，网络结构表征学习和推荐目标在模型层面上的有机统一也是一个难点；而跨领域推荐对辅

助域信息挖掘不充分，迁移的知识种类较为单一，模型扩展能力相对不足。

因此，我们考虑采用融合两种方案对算法进行改进。对基于异质信息网络的推荐算法，跨领域信息的引入能够从根本上实现推荐效果对于目标域数据丰富性和稠密度的依赖；与域内算法的结合，也能弥补跨领域推荐迁移知识种类单一的不足。

本文的主要研究内容是采用基于异质信息网络的跨领域推荐融合算法，解决个性化推荐中，由于用户行为数据稀疏造成的冷启动问题，从而改善推荐效果。该融合算法通过结合 HIN 和跨领域推荐，综合异质信息和跨领域等辅助信息，更全面的挖掘可用数据在个性化推荐问题中的价值。同时，由于推荐问题中，评分预测和 Top-K 列表推荐两类场景间存在需求和目标上的差异，因此针对两种场景分别对跨领域异质信息的挖掘进行研究。

本文工作的难点在于：

(1) 设计有效的融合策略。基于异质信息网络的推荐和跨领域推荐进行融合的难点在于，如何找到可行且有效的方法将两种算法进行有机结合。要把网络作为系统中数据和信息的载体，就必须找到合适的桥梁，将来源于目标域和辅助域的信息引入同一个数据空间中，同时保证用于知识迁移的桥梁能够尽可能多地传递有效的辅助信息，以最大限度发挥跨领域推荐的作用。

(2) 获得有效的网络表征学习。基于异质信息网络的推荐核心在于，网络节点的嵌入式表达中是否能地包含有效的用户和项目的个性化特征。尤其是在引入了辅助域信息之后，数据类型更加复杂。如何进行有效的异质信息挖掘，同时对源域和目标域信息进行平衡，避免负迁移，都是在获得表征学习中需要解决的重点问题。

(3)实现具有挖掘异质信息能力的端到端推荐模型。如 1.2.2 节所述，为了有效地利用复杂多样的异质信息，很多工作在网络拓扑层面，基于 HIN 进行网络表达学习，训练网络节点的向量。但这样的节点向量包含的网络拓扑信息不直接适用于推荐。一般情况下，节点在网络空间中的距离和真实的偏好匹配程度间存在差异[7]。所以基于网络拓扑的节点表达往往只能作为中间量作用于另外的以个性化推荐为目标模型。因此，如何实现具有挖掘异质信息能力的端到端模型，使训练获得的节点表达，既包含异质信息网络的拓扑特征，又能同时挖掘到用于个性化推荐的偏好信息，是本文工作的另一个难点。

通过大量的工作，本文解决了上述问题。本文的主要贡献为：

(1) 针对个性化推荐中用户行为数据缺失的问题，提出引入异质信息和跨领域信息作为辅助信息来提高推荐效果。采用标签作为桥梁，构造跨领域异质信息网络，将基于异质信息网络的推荐和跨领域推荐进行有机结合，实现推荐算法的融合。

(2) 针对评分预测推荐场景，提出基于网络表达的融合算法框架 HecRec 以挖掘复杂多样的跨领域异质信息，同时采用了“立交桥式”向量处理方法，避免了信

息间冲突造成知识的负迁移情况，以最大化融合框架的优势提高推荐效果。最终融合算法框架的绝对平均误差为 0.6384，比相关工作减少了 2.7%。

(3) 针对 Top-K 列表推荐场景，提出基于嵌入传播层的融合算法模型进行跨领域异质信息的挖掘，并在端到端模型中实现推荐效果的提升。融合模型在真实数据集上的准确率、召回率和归一化折损累计增益分别达到了 0.13、0.16 和 0.22，和相关工作相比分别提升了 2.8%、1.2%和 3.0%。

1.4 论文组织结构

本文整体的组织结构如下：

第二章为本文研究工作中相关的知识背景和技术介绍。包括网络表达的常见相关算法，基于网络的推荐的相关策略，跨领域推荐有代表性的工作，以及推荐常用的评估指标等。

第三章提出了评分预测场景下的、基于网络表达的融合推荐框架 HecRec。我们通过数据观测和特征分析确定融合策略，设计适应跨领域场景的基于元路径的异质信息表达算法，并结合推荐目标，设计了基于异质信息网络表达的跨领域推荐框架。通过对比实验，对其推荐性能进行了详细评估。

第四章提出了 Top-K 列表推荐场景下的、基于嵌入传播层的融合推荐模型。该模型通过基于嵌入传播层的神经网络结构对跨领域异质信息进行有效的挖掘，同时基于推荐目标对嵌入式向量进行训练，实现了端到端的模型结构。最后通过具体实验，对模型的优越性进行了具体验证。

第五章对全文内容进行了总结。阐述了本文介绍的两方面工作的主要内容和贡献，并对未来进一步的研究方向进行了介绍。

2 技术背景

本章介绍研究工作相关的技术背景，包括传统推荐算法、网络表达相关算法和跨领域推荐经典算法。同时介绍推荐系统效果的评估方法以及实验中用到的开发平台。

2.1 传统推荐算法

本节内容主要介绍传统推荐的相关概念和常用算法。如 1.2.1 节所述，中介绍传统推荐算法中实现个性化的方法主要有两类，分别是基于内容和协同过滤。不同的推荐策略过程和依据的假设有明显差异。不同算法虽然在实现上各有差异，但是算法的预测能力都依赖于相似性度量的选择来评估用户和项目之间的相似程度。下面会详细介绍推荐领域常用的相似性度量。另外作为协同过滤的经典算法，对基于评分矩阵分解的个性化推荐算法也将进行详细介绍。

（1）常用的相似性度量

在推荐领域中，主要有以下几种常用的相似性度量方法。为了方便表达，用X和Y表示用户x和y的n维评分向量或其他特征表示向量。

欧氏距离：许多相似度指标的基础都是欧氏距离。向量X和Y的欧氏距离定义为：

$$d(X,Y)=\sqrt{\sum_i^n(X_i-Y_i)^2} \tag{2-1}$$

欧氏距离是两个向量的相应元素之间的平方差之和的平方根。欧氏距离仅适用于以相同比例测量的数据。距离越大，两个用户或项目间的相似度越低；相反，则两个用户或项目间的相似度越高。

皮尔逊相关系数：和欧几里得距离不同，皮尔逊系数测量两个变量的相关性，相关性范围从-1 到 1，系数绝对值越靠近 1，相关性越强。

$$\rho(X,\ Y)\ =\ \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sum_i^n(X_i-\mu_X)^2}\sqrt{\sum_i^n(Y_i-\mu_Y)^2}} \tag{2-2}$$

余弦相似度：作为一种较为常见的相似度度量方法，余弦相似度通过计算向量间的余弦值对向量的相似性进行评估。相似度数值范围从 0 到 1，值越大，相似度越高。

$$C\left(X,\ Y\right)=\frac{\sum_i^n X_i\times Y_i}{\sqrt{\sum_i^n(X_i)^2}\sqrt{\sum_i^n(Y_i)^2}} \tag{2-3}$$

在推荐场景下，可以通过不同用户对项目的评分向量的余弦相似度来判断用

户对同一批项目是否有相似的偏好。

杰卡德相似度（Jaccard Similarity）：是由 Paul Jaccard 提出的，最初用于度量集合之间的相似度指标。其定义为两个集合交集大小与并集大小之间的比例。杰卡德相似度的数值范围大于 0 小于 1。

$$J(X,\ Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \tag{2-4}$$

（2）基于矩阵分析的推荐算法

作为推荐领域的经典算法，基于矩阵分解的推荐被广泛应用到各种推荐平台。相比较传统的基于相似度协同过滤算法，矩阵分解能够更好的捕捉用户的偏好和项目特征以及用户项目间的匹配关系。同时，大量用户和项目的相似度计算的避免也在时间复杂度上进行了优化。简单来说，基础的 MF 算法的核心思想在于，分别将用户和项目映射到联合的特征空间中，每个用户或项目获得其对应的特征向量，然后用向量的内积拟合用户和项目间的交互行为。对于任意用户u，对应一个n维向量X_u来表示该用户的个性化特征，向量中的元素表示用户对某种特定特征的偏好情况。对于任意项目i，对应一个n维向量Y_i，向量中的元素表示项目的某种特定特征的显著情况。向量间点积的结果${X_u}^{\mathrm{T}}Y_i$表示用户对项目的评分情况，即：

$$\hat{R}_{u,i} = {X_u}^{\mathrm{T}}Y_i \tag{2-5}$$

用户和项目特征向量学习的过程，就是通过拟合行为数据进行向量修正的过程。常用的目标函数为加正则的平方误差损失函数（以评分数据为例）：

$$\min_{X,\ Y} \sum_{(u,\ i)\epsilon \mathrm{K}} \left(R_{u,\ i} - {X_u}^{\mathrm{T}}Y_i\right)^2 + \lambda(\|X_u\|^2 + \|Y_i\|^2) \tag{2-6}$$

其中K是训练集中的用户-项目对集合，λ是正则项系数。算法通过拟合观察到的用户评分来学习模型，目的是通过预测未知评分的方式来提取历史评分数据所包含的用户项目特征并进行推荐。参数λ用来控制正则化的程度，通常由交叉验证等方式确定。

（3）不同推荐场景下的优化算法

对于上文中提到的两种推荐场景，评分预测和Top-K列表推荐，由于它们分别针对不同的优化目标，因此也对应不同的训练算法。

a) 评分预测场景

评分预测场景中，以具有物理意义的真实评分为目标。因此，相关的推荐模型往往是基于单点法（pointwise）的模型，训练过程中的输入单位为单个样本的形式，每个样本对应一个预期结果。模型本质上是把个性化因素作为特征的回归问题。因此回归问题相关的优化思想都可以应用于评分预测类的推荐模型。常用的目标函数可以是平方误差损失函数等。本节中介绍的基于矩阵分解的推荐算法中，采用式

（2-6）对模型进行优化训练的过程就是pointwise思想，保证获得的特征向量能通过内积计算最大程度的接近实际用户评分。

b) Top-K列表推荐场景

在列表推荐场景下，由于缺乏评分等显式反馈数据，往往采取pairwise算法对模型进行训练，即训练过程的输入单位为成对样本。在推荐领域中，最常用的pairwise算法为贝叶斯个性化排序（BPR）算法。

在BPR算法中，对于任意一个用户u，存在该用户喜欢和不喜欢两个项目集合，即用户u的正向和负向相关项目。对于两个集合中分别取出的项目i和j，就可以获得一条用户训练的三元组样本<u,i,j>。这表示，最终学习到的特征向量，需要满足用户u对于i的评价排序要比j靠前的约束条件。目标函数可以表示为：

$$Loss = \sum_{(u,i,j)} -ln\sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right) + reg \tag{2-7}$$

其中reg为正则项，(u,i,j)为用户对应的三元组。通过最大化正向反馈项目和负向反馈项目之间的排序差距，对项目进行序列分类。具体排序分值$\hat{y}_{ui}$如何计算则由模型本身决定。以矩阵分解为例，则采用内积对用户和项目间的关系进行建模，计算排序分值。

2.2 异质信息网络

异质信息网络作为推荐领域的新兴研究方向，能较好的挖掘类别多样的异质信息和复杂关系。本节内容将主要介绍基于异质信息网络的推荐算法的常用相关概念。

异质信息网络（HIN）。异质信息网络是一种特殊的复杂网络结构。网络结构可以由$\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$的形式表达，其中$\mathcal{R}$表示网络中连边的集合，$\mathcal{V}$表示网络中的节点集合。节点和连边构成一个基本的网络结构。而异质信息网络和普通的网络相比的特殊之处在于，节点和边的类型更加复杂。如果网络中每个节点v和连边r都满足一个映射关系$\varphi(v) \to T_v$和$\varphi(r) \to T_r$，其中T_v和T_r分别表示节点和连边的类型，则任意异质信息网络满足：

$$|T_v| + |T_r| > 2 \tag{2-8}$$

即节点和边的种类之和大于两种。不满足条件(2-8)的普通网络结构，即节点和连边不存在不同的类型的网络结构，也被称为同质（Homogeneous）信息网络。如图 2-1 所示，在电影推荐系统中，存在如下四种可能出现的节点类型，用户（User，U）、项目（Movie，M）、明星（Star，S）和电影类别（Genre，G）。不同类别的节点之间存在多样的互动、偏好或匹配关系。用户与项目的互动行为直接体现用户的项目

偏好，而用户和其他节点的关系也包含了用户的个性化信息，从异质信息网络的结构中可以看出异质信息对于个性化特征挖掘具有重要价值。

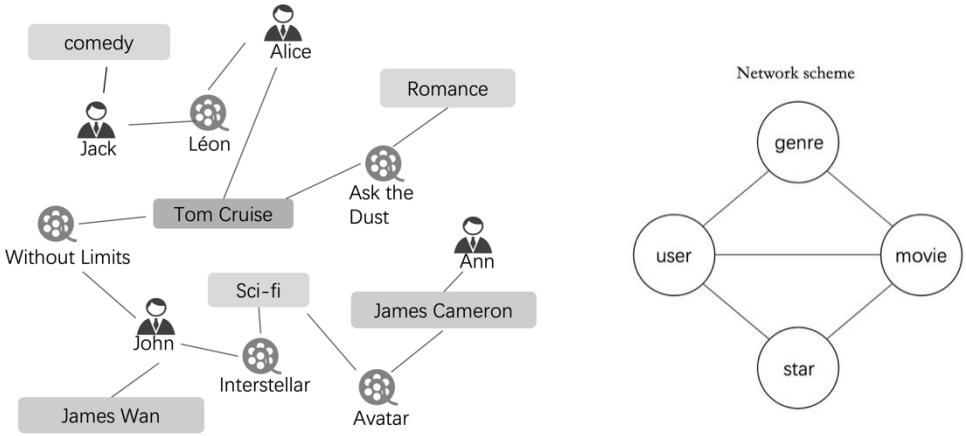

图 2-1 电影推荐系统中的异质信息网络

Figure 2-1 A prototype of a HIN in a movie recommendation system

元路径（Meta-path）。异质信息网络具有多类别节点的特征，不同类别的节点需要进行区别处理。为了便于区分不同类型节点的物理意义，更好的挖掘网络中的异质信息，一些工作中开始引入了元路径这一概念。元路径（Meta-path）是节点类型及节点间边的类型组成的复合关系序列。一条元路径ρ可以表示为$\mathcal{V}_1 \xrightarrow{\mathcal{R}_1} \mathcal{V}_2 \xrightarrow{\mathcal{R}_2} \cdots \xrightarrow{\mathcal{R}_l} \mathcal{V}_{l+1}$的形式，它用于描述$\mathcal{V}_1$和$\mathcal{V}_{l+1}$之间的一种复合连接关系$\mathcal{R}_1 \circ \mathcal{R}_2 \circ \cdots \circ \mathcal{R}_{l+1}$，其中$\mathcal{V}_n$表示第n个节点类型，$\mathcal{R}_n$表示第n个连边类型，$\circ$表示连边类型间的复合操作。在图 2-1 中，元路径$User - Star - Movie$可以表示用户和电影之间基于明星类型的相互匹配关系，基于这条元路径可以获得节点序列“Ann-James Cameron-Avatar”，进而得知用户“Ann”可能对电影“Avatar”感兴趣。基于元路径把网络结构转化为节点序列的形式进行分析，有利于更好地捕捉网络中的异质信息和节点间的复杂关系。现有很多基于异质信息网络的工作都利用的元路径的特点和优势进行。

随机游走（Random walk）。随机游走是指在网络结构上生成节点序列的动态过程。在网络结构上给定一个起始节点作为当前节点，从当前节点的邻居节点中随机选择一个节点，作为下一跳的当前节点，如此循环下去直到序列到达指定长度，获得的节点序列即由随机游走生成的路径（Path）。对于网络结构$\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$，考虑一次随机游走过程：从节点$v_1$开始，第 t 步的当前节点为v_t，则下一步选择节点v_x的概率为：

$$\mathrm{P}(n_{t+1} = v_x | n_t = v_t) = \begin{cases} \left|\frac{1}{\mathcal{N}(v_t)}\right|, & |\mathcal{N}(v_t)| > 0 \\ 0, otherwise \end{cases} \tag{2-9}$$

其中，n_{t+1}表示序列中第$t+1$个节点，$\mathcal{N}(v_t)$表示节点v_t的邻居节点集合，也就是说每轮选择下一跳节点时，邻居节点等概率随机选择，反复重复单轮操作，在节点序列达到预设长度时候，获得一次随机游走对应的节点路径。此时，不同类型的节点不进行类别上的区分，即同质网络的随机游走。

而在异质信息网络上，随机游走的节点选择，不能直接将所有类型节点统一处理，这样会损失异质信息，无意义路径的产生也会一定程度上引入噪声。因此 HIN 中的随机游走，往往基于元路径实现。根据元路径ρ的复合策略$\mathcal{V}_1 \xrightarrow{\mathcal{R}_1} \mathcal{V}_2 \xrightarrow{\mathcal{R}_2} \cdots \xrightarrow{\mathcal{R}_l} \mathcal{V}_{l+1}$，第$t+1$个节点的选择概率为：

$$P(n_{t+1} = v_x | n_t = v_t, \rho) = \begin{cases} \left|\frac{1}{\mathcal{N}^{\mathcal{V}_{t+1}}(v_t)}\right|, & \left|\mathcal{N}^{\mathcal{V}_{t+1}}(v_t)\right| > 0 \text{ and } \varphi(v_x) = \mathcal{V}_{t+1} \\ 0, otherwise \end{cases} \tag{2-10}$$

其中，$\mathcal{N}^{\mathcal{V}_{t+1}}(v_t)$表示节点$v_t$的邻居中，属于类别$\mathcal{V}_{t+1}$的节点的集合。在 HIN 中进行随机游走，节点选择策略是有元路径策略指导的，最终获得的节点序列满足对应元路径的节点类型复合关系的描述。元路径的设计可以灵活地赋予节点序列以有效的物理意义，因此基于元路径的随机游走可以更好的挖掘 HIN 中有效的异质信息。

2.3 网络的嵌入表达算法

推荐问题，最根本的目标在于，如何更加全面的了解用户的个性化偏好和项目的特征。而这些内容除了由直接行为体现，如用户对项目的评分及项目本身的属性描述等，还可以通过间接关系获得。比如给同一个项目打高分的两个用户，更有可能具有相似的偏好；具有多个共同的好评用户的项目，有可能具有相似的目标用户群。因此提高推荐效果，要更充分的挖掘数据中包含的非直接互动信息。

网络表达旨在将网络结构信息转化为连续的特征值，用低维向量来表示网络中的节点。网络表达过程中的关键是保留网络中的拓扑关系，识别网络中即使没有直接但具有相似特性的节点。本章内容主要介绍推荐领域常用的网络表达相关算法。

2.3.1 同质网络的表达算法

作为网络表达算法的基础，同质网络学习算法主要从结构上分析节点的相似

度和差异性，表达目标较为清晰。同质网络节点类型单一，如 2.2 节中介绍，网络结构相对简单，网络节点类型没有区分，不同节点间的差别不体现在节点本身，而体现在以节点为中心的局部网络结构的差别。常用的同质网络表达算法主要有以下几种。

（1）深度游走算法（DeepWalk）

在深度游走[34]算法中，无监督深度特征学习第一次被用在网络表达算法中来学习网络结构的节点的低维特征向量。算法结合了随机游走和深度语言学习模型 SkipGram 来获得节点的低维嵌入式特征向量。深度游走算法实际上是一种特殊的神经语言模型，可以将节点序列作为特殊形式的语言进行处理。自然语言处理模型可以捕捉语义和句法信息，基于网络的深度游走算法中，信息则更多隐藏在的在网络拓扑结构中。

深度游走算法主要考虑的问题是如何将一个同质信息网络，如社交网络中的节点进行向量化表达的问题。对于网络结构$\mathcal{G} = \{\mathcal{V}, \mathcal{R}\}$，深度游走的目标就是学习节点的嵌入式特征向量$X_E \epsilon \mathbb{R}^{|\mathcal{V}| \times d}$来对节点的隐式特征进行描述，其中$d$是一个相对网络规模较小的向量维度值。

DeepWalk 算法主要包括两个主要阶段，第一个阶段通过随机游走获得节点序列，第二个阶段基于节点序列进行参数训练从而获得节点的向量化表达。同质信息网络中的随机游走的过程如 2.2 节中所述，设定路径的长度上限，变换起始节点，循环进行随机游走的过程，生成一系列节点序列。受 word2vec 的启发，在 DeepWalk 算法中，随机游走生成的节点序列作为语料信息中的句子，采用自然语言处理模型 SkipGram 算法进行参数更新训练向量的过程。该算法首先对句子进行窗口划分，窗口内一起出现的单词对作为相关性高的二元组输入，通过最大化二元组中单词共同出现的概率来训练单词的向量。在深度游走算法中，将节点序列用同样的方法进行处理，来获得基于网络结构相似性的节点的向量化表示。其优化目标公式化的结果如下所示：

$$\max_{f} \sum_{v \epsilon \mathcal{V}} log\ Pr\big(\mathcal{N}(v) \big| f(v)\big) \tag{2-11}$$

其中f表示从从节点到嵌入式表达的映射关系，模型训练目标是如果当前节点为v，下一个节点，即共同出现的节点为v在网络中实际的邻居节点的概率最大化。即网络结构中，节点的距离越近，向量表达的相似度越高。

（2）大规模网络信息嵌入算法（LINE）

LINE[18]算法是一种适应范围较广的网络结构表达算法。除了上述结构较为简单的同质无向网络外，信息网络结构也可以是有向网络，如学术引用网络等。网络中的连边的具有不同形式的权重，可以是二值化权值也可以是连续值。对信息网络进行嵌入式表达学习，必须保留不同类型网络中的信息。

在网络结构中，一阶邻接关系即有连边的节点。w_{uv}表示节点u和v之间的一阶邻接权重。二阶邻接关系可以表达为p_u和p_v间的相似度，其中$p_u = (w_{u1}, w_{u2} \cdots w_{u|\mathcal{V}|})$，即$u$和网络中所有节点间的一阶权重向量。在 LINE 算法中，同时保留一阶和两阶邻接关系以挖掘网络结构信息。一阶邻接关系指网络中直接相连的两点间的关系，对于无向连边(i,j)连接的两个节点v_i和v_j，定义联合概率为：

$$\mathrm{p}_1(v_i, v_j) = 1 \,/\, exp(-e_i^T \cdot e_j) \tag{2-12}$$

其中$e_i \epsilon \mathbb{R}^d$是节点$v_i$的低维特征表达，同时定义经验分布为$\widehat{p_1}(v_i, v_j) = w_{ij} \,/\, W$，其中$W = \sum_{(i,\mathrm{j})\epsilon\mathcal{R}} w_{ij}$，优化目标为最小化联合概率和经验值的距离，即：

$$O_1 = d(p_1, \widehat{p_1}) = -\sum_{(i,\mathrm{j})\epsilon\mathcal{R}} w_{ij}\, log p_1(v_i, v_j) \tag{2-13}$$

其中$d(\cdot,\cdot)$是衡量两个分布的距离的常用函数 K-L 散度。一阶连通相似性适用于无向网络结构中。而对于有向网络，每条连边是从头结点指向尾节点。在有向图中的每个节点，维护两个嵌入式向量，一个用来表示表示该节点本身特征，另一个向量是该节点作为其他节点的上下文邻居节点时的特征向量。对于一对相邻节点v_i和v_j，给定节点v_i的情况下，产生上下文邻居节点v_j的概率为：

$$p_2(v_j|v_i) = \frac{exp(e_i^T \cdot e_j)}{\sum_{k=1}^{|\mathcal{V}|} exp(e_i^T \cdot e_k)} \tag{2-14}$$

经验分布概率：

$$\widehat{\mathrm{p}_2}(v_j|v_i) = \frac{w_{ij}}{\sum_{k=1}^{|\mathcal{N}(v_i)|} w_{ik}} \tag{2-15}$$

目标函数为基于 K-L 散度的相似度度量：

$$O_1 = d(p_2, \widehat{p_2}) = -\sum_{(i,\mathrm{j})\epsilon\mathcal{R}} w_{ij} \log p_2(v_j|v_i) \tag{2-16}$$

和 DeepWalk 算法相比，LINE 算法更接近于宽度优先搜索的思想，而 DeepWalk 的随机游走体现的则是深度优先的思想。他们的主要区别在于对节点的相似度定义。

（3）Node2vec

Node2vec[10]算法被为是 DeepWalk 算法的扩展，它提出了有偏置的随机游走，同时结合了深度优先和广度优先两种思想，也就是在勘探和开采之间进行平衡（Exploitation-exploration trade-off）。对于网络中的节点，通过邻居节点采样策略获得相邻节点集合，用$\mathcal{N}_s(v_i)$表示通过采样策略获得的节点v_i的邻居节点，然后采用有偏置的随机游走[10]，即下一跳节点选择时非等概率的随机游走算法，结合 SkipGram 算法对节点向量进行训练。类似式(2-11)所示，Node2vec 算法的优化目标为：

$$\max_f \sum_{v\epsilon\mathcal{V}} log\ Pr(\mathcal{N}_s(v)|f(v)) \tag{2-17}$$

为了使目标函数可解，算法中提出两个重要假设：首先假设给定已知节点，该节点的不同的邻居节点出现在上下文中的概率分布是相互独立的；另外，一个节点作为源定点和邻接节点共享相同的嵌入式表达，这是区别于 LINE 算法的比较重要

的基础假设。基于上面的假设，可以获得以下公式描述：

$$Pr(\mathcal{N}_s(v_i)|v_i) = \prod_{v_j \in \mathcal{N}_s(v_i)} Pr\left(v_j \middle| v_i\right) \tag{2-18}$$

$$Pr\left(v_j \middle| v_i\right) = \frac{exp(e_i^T \cdot e_j)}{\sum_{k=1}^{|\mathcal{V}|} exp(e_i^T \cdot e_k)} \tag{2-19}$$

基于以上假设及公式描述，Node2vec 的目标函数为：

$$\max_f \sum_{v \epsilon \mathcal{V}} \left(-log\ Z_{v_i} + \sum_{k=1}^{|\mathcal{N}_s(v_i)|} e_i^T \cdot e_j\right) \tag{2-20}$$

其中归一化因子：

$$Z_{v_i} = \sum_{k=1}^{|\mathcal{N}_s(v_i)|} exp(e_i^T \cdot e_k) \tag{2-21}$$

该计算方式的复杂度较高，所以采用负采样策略对邻居节点进行采样，具体内容可参考文献[10]。

2.3.2 异质网络的表达算法

上一节中介绍到的算法（DeepWalk、LINE 和 Node2ve）c 都是较为经典的基于同质网络结构的表达算法。显然这些算法不能很好地适应于异质信息网络。最主要的原因在于异质信息网络中节点类型和连边类型复杂，不同类型的节点间需要结合节点本身含义采用不同的方式处理，同质网络表达算法会导致由节点和连边类型携带的信息损失。另外同质信息网络中节点差别主要在于局部网络结构的不同，而实际上在 HIN 中，节点间连边本身也包含特定的语义信息。因此异质信息挖掘需要更适用于异质信息网络的嵌入式表达算法来完成。

（1）Embedding of embedding（EOE）

EOE[35]算法将异质信息网络的表达问题转化为同质网络来解决。对于复杂度较低的异质信息网络，EOE 算法考虑将完整网络按照节点类型分成不同的子网络，分别基于子网对节点进行嵌入式表达。算法命名为“嵌入表达的表达”是因为不同子络的训练结果在不同的特征空间，所以再通过转移矩阵将基于不同子网的表达映射到同一个特征空间中。

以存在用户和项目两种节点的网络为例，将耦合网络表示为$G_{uv}(G_u, G_v, E_{uv}, W_{uv})$，其中$G_u, G_v$表示两个同质子网，$E_{uv}$表示子网络间的连边，$W_{uv}$表示子网间连边的权重。同质网络$G_v$中，节点间相连的概率为：

$$p\left(v_i, v_j\right) = \frac{1}{1+exp(-v_i^T \cdot v_j)} \tag{2-22}$$

其中v_i和v_j为G_v中节点的嵌入式向量。对于子网间跨网络的节点对，节点间连接的概率为：

$$p\left(u_i, v_j\right) = \frac{1}{1+exp(-u_i^T M v_j)} \tag{2-23}$$

其中u_i和v_j分别为G_u和G_v中的节点表达，$M \in \mathbb{R}^{|U| \times |V|}$为转移矩阵，作为两个子网

嵌入表达空间的桥梁。对于有连边的节点对，目标函数为（以G_u为例）：

$$O_1 = -\sum_{(i,j)\in E_u} w_{ij} log\left(p(u_i, u_j)\right) \tag{2-24}$$

没有连边的节点对的目标函数为：

$$O_2 = -\sum_{(i,j)\notin E_u} log\left(1 - p(u_i, u_j)\right) \tag{2-25}$$

对于子网G_v和E_{uv}的情况类似，分别为有连边和无连边的节点对的目标函数，因此完整的目标函数共包括六部分，同时考虑同类型节点网络中的结构特征，同时考虑不同子网间连边，即异质连边的有效信息，将 HIN 的网络表达转化为同质网络表达问题来解决。但 EOE 算法仅适用于网络节点类型较少的情况，且不适用于二部图（Bi-partite graph）问题（不存在相同节点间的连边），因此适用范围较小，不能解决较复杂的异质信息网络的嵌入式表达问题。

（2）Metapath2vec

针对 DeepWalk 算法无法适用于异质信息网络的问题，Metapath2vec[36]算法提出了一种扩展的方法。

为了更好的保留异质信息网络中包含的丰富的语义信息，该算法提出了基于元路径的随机游走策略，也就是由元路径指导的有方向的节点选择替换 DeepWalk 中的随机节点选择。其中为了挖掘网络中的相似节点，元路径的设计一般具有对称特性，细节部分在 2.2 节中随机游走部分中进行了详细介绍。在图 2-1 中的异质信息网络中包含有四种节点类型，用户（U）、项目（M）、类型（G）和明星（S），以元路径$\mathrm{U}-\mathrm{M}-\mathrm{U}$为例，基于此元路径进行随机游走生成的节点路径$user1-movie2-user3\cdots-user\ n$中，$user3$作为$user1$的上下文（假设窗口长度大于 3），两个用户虽然没有在网络拓扑中直接相连，但是具有相同的项目交互历史，可以认为具有基于电影的相似偏好；类似的，元路径$\mathrm{U}-\mathrm{M}-\mathrm{S}-\mathrm{M}-\mathrm{U}$产生的节点序列中，用户和某电影明星可能并未直接相连，但是其作为电影的一项特点，将有共同点的电影，及有相同爱好的用户都挑选出，同时出现在上下文中。因此可以认为，元路径指导的随机游走可以较好的保留网络中对目标有效的异质信息。

（3）HIN2vec

HIN2vec[37]算法为了挖掘网络中的异质语义信息，提出了一种基于神经网络的异质信息网络表达算法。不同于其他基于随机游走的网络表达算法，如 DeepWalk 和 Metapath2vec，HIN2vec 算法虽然也是基于随机游走和节点序列的方法，但是它用神经网络结构替换了 SkipGram 算法进行向量训练，不仅能学习节点间相关关系，同时将异质连边的信息嵌入目标向量空间。

HIN2vec 算法框架同样也分为训练数据生成阶段（基于随机游走策略）和嵌入式表达学习阶段。数据生成阶段采用类似的随机游走和负采样策略，虽然是基于

HIN 的随机游走，但没有基于元路径进行，而是对不同类型的邻居节点进行随机选择，而是将不同的元路径类型用 one-hot 的方式体现在神经网络的标签中。即第一阶段生成的数据结构为$\langle x, \mathrm{y}, l\rangle$，其中$x$和$y$分别为节点对的两个节点，$l$为$x$和$y$间的元路径关系。这种方法不需要人为对网络特点进行分析设计元路径类型，并且降低了随机游走策略的时间成本。在向量训练阶段，输入三元结构的 one-hot 形式进入神经网络模型，通过预测x和y间有没有对应的元路径连接关系进行参数更新。

2.4 跨领域推荐技术

为了解决个性化推荐系统，跨领域推荐算法作为新兴的研究方向获得了广泛的关注。跨领域推荐主要通过引入辅助域信息来弥补目标域中用户行为信息的不足，从而提高模型的预测和推荐效果。不同领域之间是存在内在联系的。常见的推荐领域都是针对单一推荐项目的，例如 Netflix 网站主要针对电影和电视剧的视频类进行推荐，Last.fm 主要针对音乐作品进行推荐。不同用户对于不同领域商品的偏好并不是完全独立且不相关的。比如用户对于视频和音乐的品味相互之间也存在影响，文艺片爱好者和恐怖片爱好者相比，前者喜欢古典音乐的概率整体上看相对较高。因此在一个领域中获得的用户品味偏好信息等相关知识可以迁移到其他领域适应性地进行利用。这种相关领域间的内在联系就是跨领域推荐技术发挥作用的必要基础。

对于跨领域中“域”的认识，在相关的工作中并没有一个统一的标准。有的工作中，重叠的用户发挥重要作用，所以将同一个推荐系统中不同的项目类型作为不同的领域[38]，如电商系统中各种不同类别的商品；有些工作[39, 40, 41, 52]将同一类项目的数据集进行拆分获。如电影推荐数据中，根据不同的流派，将数据集进行拆分。针对我们要解决的问题，在我们的工作中，将“域”（domain）定义为用户和项目没有必要相关性的，具有不同类型推荐项目的主题推荐平台或推荐系统。这样选择的原因主要是，对于解决冷启动问题，需要在用户行为活跃程度有偏差的环境下的方案才有意义，而参考文献[39, 40, 41]等工作中的处理方案实际上并不适用；另外，对于领域的项目类型，越相似的场景对于跨领域的难度就会越低，在项目类型不同但相关的领域间探索的跨领域方案实际上可用范围更广。

目前现有的跨领域推荐算法可以分为两类[42]，一种策略着重于提取不同领域间的共享信息，主要针对某种特定的共享信息，进行抽象和迁移；另一种策略是整合和利用分布在不同系统中用户偏好相关的特征，力求模型具有较好的泛化能力。

（1）Codebook Transfer

作为跨领域推荐中较为经典的算法，CBT[32]采用迁移学习的方式解决协同过

滤算法中的数据稀疏问题，该算法通过对评分矩阵的联合聚类操作，提取用户群组的评分模式，成为密码本（codebook），并将辅助域提取的评分模式应用到目标域以预测目标域的缺失评分。首先利用 ONMTF[43]对辅助域评分矩阵X_a进行分解，得到用户和项目包含聚类因子的矩阵 U 和 V，目标函数为：

$$\min_{U\geq 0,V\geq 0,S\geq 0,} \|X_a - U_a S V_a^T\|^2 \tag{2-26}$$

其中 U 和 V 都是单位正交矩阵，每一行都只有一个非负值来表示聚类结果，通过 U 和 V 及辅助域的评分矩阵可以构造密码本：

$$B = [U_a^T X_a V_a] \oslash [U_a^T I I^T V_a] \tag{2-27}$$

其中⊘表示点除运算，式(2-27)表示将用户-项目聚类群组的评分进行平均，获得二维群组的评分模式（the cluster-level rating pattern）。在获得了从辅助域提取出的评分模式信息 B 之后，目标域的评分矩阵X_t可以通过密码本的展开来进行预测，即缺失部分的评分通过复制辅助域中对应群组的用户对项目的评分来进行填充。首先需要将目标域中的用户和项目按照辅助域的聚类结果进行聚类，也就是为目标域中的用户和项目找到匹配的群组：

$$\min_{U_t,V_t} \|(X_t - U_t B V_t^T) \circ W\|_F^2 \tag{2-28}$$

其中U_t和V_t是目标域用户和项目应该匹配的聚类群组矩阵，矩阵的每行每列只有一个值为 1，其余为 0，寻找聚类结果使得和现有的目标域评分矩阵的差距最小。W为掩码矩阵，将X_t中有评分的部分提取出来，即有评分的元素为 1，没有为 0，∘为矩阵的点对点乘积运算。目标函数的优化算法详细过程见[32]。目标域评分矩阵的填充结果计算公式为：

$$\widetilde{X_t} = W \circ X_t + (1 - \mathrm{W}) \circ (U_t B V_t^T) \tag{2-29}$$

也就是把原有评分和填充评分矩阵进行合并的过程。

（2）TagCDCF

基于信息整合的跨领域推荐算法需要将不同领域中的有效特征提取出来，包括显式信息——用户和项目属性等，以及隐式特征。因为相对于整个系统的规模，某一个用户或项目的直接相关信息较少，因此往往隐式特征中包含的信息更加丰富。要想整合信息，最核心的难点在于，如何将不同特征空间的有效信息映射到同一个共享空间中。

Yue 等人[29]提出了 TagCDCF 算法，即基于标签的跨领域协同过滤算法。不同领域间，以标签作为桥梁进行不同领域间的联合，以解决协同过滤冷启动。算法首先定义不同领域间基于公共标签的相似度：

$$S_{ip}^{(U)} = \frac{\sum_{t\in CT}\left(A_{ix(t)}^{(1)} A_{py(t)}^{(2)}\right)}{\sqrt{\sum_{t\in CT}\left(A_{ix(t)}^{(1)}\right)^2}\sqrt{\sum_{t\in CT}\left(A_{py(t)}^{(2)}\right)^2}} \tag{2-30}$$

$S_{ip}^{(U)}$表示一个领域中的用户i和另一个领域中用户p之间基于公共标签CT的相似度，

其中$A^{(k)}$表示第k个领域内的用户-标签矩阵，如果第i个用户使用过第l个标签，则$A_{il}^{(k)}$为 1，否则为 0。式(2-30)实际为用户间基于公共标签的余弦相似度，同理项目间相似度为：

$$S_{jq}^{(V)}=\frac{\sum_{t\in CT}\left(B_{jx(t)}^{(1)}B_{qy(t)}^{(2)}\right)}{\sqrt{\sum_{t\in CT}\left(B_{jx(t)}^{(1)}\right)^{2}}\sqrt{\sum_{t\in CT}\left(B_{qy(t)}^{(2)}\right)^{2}}} \tag{2-31}$$

然后将经典单领域协同过滤 PMF 算法[43]进行跨域场景的适应性扩展，引入跨域的用户和项目相似度，通过最小化预测评分和真实评分的距离训练用户和项目嵌入式向量[33]。

2.5 推荐效果评估方法

系统的评估方案随推荐场景的不同有所区别。在评分预测场景中，算法的目标是对用户的个性化评分进行预测，即预测结果为一定范围内的连续数值型标签，所以推荐系统实质上为回归问题，评估推荐效果可以采用回归问题的常用评估指标如 MAE 和 RMSE。而在 Top-K 列表推荐场景中，系统的目标是识别出用户的偏好项目，这时系统解决的问题实质上是标签不均衡的分类问题。此时，生成的列表中实际被用户喜欢的项目的概率越高，模型效果就越好。这种情况下，常用指标有准确率（Precision）、召回率（Recall）、命中率（Hit）和 F1 值。还有一些工作侧重评估推荐列表的排序质量，采用归一化折损累计增益（NDCG）等指标。不同的个性化推荐场景下的推荐系统具有不同的目标，因此也对应不同的模型效果评估指标。

（1）均方根误差和绝对平均误差

作为回归问题中常用的模型评估指标，均方根误差（RMSE）和绝对平均误差（MAE）被广泛用于评分预测模型效果的评估。计算公式分别如下所示：

$$MAE=\frac{1}{|R|}\sum_{(u,i)\in R}|r_{ui}-\hat{r}_{ui}| \tag{2-32}$$

$$RMSE=\sqrt{\frac{1}{|R|}\sum_{(u,i)\in R}(r_{ui}-\hat{r}_{ui})^{2}} \tag{2-33}$$

其中 R 为测试集的评分集合，r_{ui}为测试集中某用户 u 对项目 i 的真实评分值，$\hat{r}_{ui}$为模型的预测评分。MAE 是绝对误差的平均值，能更直观的表现误差的情况，RMSE 是预测值与真实值偏差的均方根，和 MAE 相比，更能反映误差的方差情况，受异常值的影响更大。

（2）准确率、召回率和 AUC

我们首先基于二分类问题介绍相关评估指标的基本概念，之后结合推荐场景介绍相关指标的具体计算公式。

要理解分类问题评估指标，就需要先了解混淆矩阵的含义。二分类问题中，测

试集样本本身具有真实类别$y \in \{0,1\}$，分为正例和负例；模型需要对测试集的样本集合，进行类别预测，模型获得预测类别$\hat{y}_i \in \{0,1\}$。如图 2-2 所示，混淆矩阵中是用来评估分类器结果和真实情况的常用相关指标。其中 TP 代表的是预测为正的样例中实际为正的样本，FN 是预测为负的样例中实际为正的样本，FP 是预测为正的样例中实际为负的样本，FN 是预测为负的样例中实际为负的样本。四部分样例相加总和为测试集中的样例总和。准确率、召回率的计算公式分别式（2-34）和（2-35）所示：

$$Precision = \frac{TP}{TP+FP} \tag{2-34}$$

$$Recall = \frac{TP}{TP+FN} \tag{2-35}$$

		预测类别	
		1	0
真实类别	1	True positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

图 2-2 混淆矩阵

Figure 2-2 The confusion matrix

准确率表示预测为正的样例中，实际为正的样本比例，着重于预测的准确程度；召回率表示真实为正的样例中，被预测出的样本比例，是一个关于覆盖率的度量。这两个指标分别表示模型不同方面的能力，但不同的模型在准确率和召回率上可能出现互相矛盾的现象，无法评判模型整体优劣，这时需要F_1值对模型进行整体评估。

除了精准度和召回率，ROC 曲线和 AUC 也常用作评估指标，对分类模型进行整体评估。ROC 曲线的横纵坐标分别是假正率（False Positive Rate，FPR）和真正率（True Positive Rate, TPR）。AUC 具体是指 ROC 曲线的线下面积，ROC 曲线上每一个点代表分类器在特定阈值下的一组 FPR 和 TPR。我们希望分类器效果尽可能准确，也就是说，FPR 要尽量小，TPR 要尽量大。当分类器为随机分类时，ROC 曲线应该是形如 y=x 的曲线，在此基础上，分类器效果越好，曲线就应越靠近坐标轴上的（0，1）点。理想情况下，曲线经过（0, 1）点，此时 AUC 最大为 1。

在 Top-K 推荐场景下，推荐系统的目标就不再是对连续性评分数值的预测了。推荐系统更为常见的工作模式为，根据用户的偏好分析，基于一定的排序标准，生

成一个推荐列表。评价推荐系统的效果如何，需要统计列表中真实出现在测试集的收到用户正向行为反馈的情况。对于测试集中任意一个用户 u，存在一个对应的项目集合I_u，是针对该用户的目标推荐向量，推荐系统生成的对用户 u 的推荐列表用$\hat{I}_u$表示。那么对于用户 u，精准率和召回率分别为：

$$Pre_u = \frac{|\hat{I}_u \cap I_u|}{|\hat{I}_u|} \tag{2-36}$$

$$Rec_u = \frac{|\hat{I}_u \cap I_u|}{|I_u|} \tag{2-37}$$

测试集整体的精准率和召回率通常取用户维度的均值。

（3）归一化折损累计增益

归一化折损累计增益（Normalized Discounted cumulative gain，NDCG）用来评价排序的质量。其中 CG（Cumulative Gain），即累计增益，表示列表中所有等级结果对应的得分总和，DCG（Discounted Cumulative Gain）的思想是，如果得分高的排在后面，那么统计得分时就应该被折损。比如一次搜索获得 5 个结果，每个结果按排序获得的评分分别是 3、2、1、3 和 2，那么DCG=3+1+1.26+1.5+0.86。NDCG 是一个相对值，是当前排序和理想排序的 IDCG（ideal DCG）的比例：

$$NDCG = \frac{DCG}{IDCG} \tag{2-38}$$

上述例子中，理想排序的评分制应为 3、3、2、2 和 1，即 IDCG=3+ 3+1.26+1+0.43。在推荐系统中，NDCG 的计算是将用户实际的打分序列排序作为 IDCG 的计算依据，用对应的预测评分序列计算 DCG，从而获得 NDCG 的值。

2.6 开发平台

本节内容主要是介绍研究工作中使用到的开发环境和集成框架，包括 Anaconda 集成环境、Scikit-learn 代码库和 TensorFlow 开发框架。

2.6.1 Anaconda 集成环境

Anaconda 是一个开源的 python 和 R 语言代码的发行版，主要用于科学计算，包括数据科学，机器学习应用，大规模数据处理和预测分析等工作，并通过软件包管理系统 conda 来简化多种软件包的安装和管理。Anaconda 也可以通过创建和管理虚拟环境，实现多版本语言共存的问题。

Anaconda 中常用的第三方软件包有以下几种。Numpy 主要用于支持大规模的数组和矩阵运算，同时也提供大量的数学函数库，引入了多维数组数据结构，提高了数组和矩阵运算的效率；Pandas 是 python 的数据分析库，提供高效能、简易使

用的资料格式 DataFrame 让使用者可以快速进行操作和资料分析；Matplotlib 和 Seaborn 是基本的可视化工具；jupyter notebook 是一轻量级的基于网页的 python 编写和运行工具，和重量级的集成 IDE 相比，具有更加方便灵活的特定，广泛用于数据分析和小程序调试。

2.6.2 Scikit-learn 算法库

Scikit-learn 是 python 的一个开源的机器学习框架，它支持有监督和无监督的多种相关算法，还提供了用于数据与处理、模型训练和模型评估的许多其他使用工具。

Scikit-learn 中包含的开源项目主要包括以下六个方面。分类算法，包括支持向量机、逻辑回归、最近邻居和随机森林等相关算法；回归算法，包括支持向量回归 SVR、脊回归和岭回归等；聚类算法，包括 K-means、谱聚类等；降维算法，包括主成分分析 PCA、LDA 算法等；用于模型选择的网格搜索、交叉验证以及 2.52 节中提到的推荐常用的一系列评估指标；数据预处理阶段的特征工程和特征提取相关算法那。本文工作中数据处理和模型评估的相关部分的代码实现均基于 Scikit-learn 实现。

2.6.3 TensorFlow 框架

Tensorflow 是适用于高效的、大规模数值计算的开源代码库。最初由谷歌开发出来，如今被泛用于各种只能计算领域。该框架中，用图表示完整的计算任务，用张量表示数据，通过变量来维护状态，在会话（session）中执行图中定义好的计算任务。一个数据流图描述了计算的过程，计算在会话被启动时开始运行。计算图的使用使基于 TensorFlow 框架运行的程序极大的优化了计算效率。一般在 python 中使用 NumPy 的 API 进行大规模数据计算，如矩阵乘法等，通常将这些计算在 python 外部环境中结合其他语言进行来提高效率，但是操作间的切换会产生开销，尤其在 GPU 和分布式环境中是，这种开销就更加不可忽视。而 TensorFlow 通过创建计算图，将整个计算过程在 python 外部进行，以减小开销，其中 python 只用于构建计算图及设计数据流方向。为提高运行效率，节省运行时间，本文的第四章的工作的代码框架主体部分就是基于 TensorFlow 实现。

2.7 本章小结

本章主要介绍了工作内容相关的技术背景。首先介绍了传统推荐算法的相关内容，然后介绍本文的背景知识，即异质信息网络的相关概念。作为基于 HIN 的推荐算法的基础，网络表达算法的相关内容在 2.3 节中进行了重点介绍，包括同质网络表达和异质信息网络表达算法。2.4 节主要介绍了跨领域推荐的经典算法。2.6 节介绍了实验中用到的开发平台和集成框架。

3 基于 HIN 表达的评分预测推荐框架

本章针对评分预测的推荐场景，提出了一种新的基于异质信息网络的跨领域融合框架，实现用户对项目的个性化评分预测。实验验证，融合算法框架的绝对平均误差为 0.6384，比现有最好的算法减少了 2.7%。

3.1 基本思想

为解决冷启动问题，本章设计融合框架 HecRec（HIN embedding based cross-domain recommendation system），以挖掘异质网络信息和跨领域信息，实现用户对项目的个性化评分预测。而采用基于网络表达的推荐算法，进行个性化评分预测，主要面临两个难点：1）如何在复杂网络结构中获取有效的网络节点表达。与异质信息网络相比，引入跨领域信息后，网络中节点类型更加丰富，网络结构更加复杂。如何设计网络表达算法对网络信息进行有效处理，使跨领域异质信息的引入实现最大的效果提升是需要解决重要问题。2）如何将网络表达算法与个性化评分预测进行有机结合。基于网络表达算法获取的节点的向量化表示，是基于网络结构信息，提取节点的网络空间信息。然而，个性化评分预测问题中所需要的用户偏好信息，并不完全等同于网络结构特征。因此，如何网络表达算法和个性化目标进行融合，如何有效利用基于网络表达算法的节点向量化表示，是需要解决的另一个难点。

为解决以上难点，本章提出采用标签作为桥梁，构造跨领域异质信息网络，然后设计特有的元路径，对复杂的网络信息提取，并提出“立交桥式”的向量处理方法，对信息进行整合和利用。同时，为了更好的将节点的向量化表达用于个性化评分预测，我们采用了扩展的矩阵分解模型训练评分预测器，对用户评分进行预测。

3.2 数据集介绍

本文工作中主要使用两个数据集进行模型效果的验证，MovieLens 和 LibraryThing。这两个数据集都是在推荐领域较多使用的公开数据集。MovieLens 数据集（ml-20m）描述了电影推荐服务网站（http://movielens.org）的 5 星评级和用户的文本标记活动。数据集中包含了 27278 部电影的 20000263 个评分和 465564 个文本标签。 LibraryThing 数据集是图书网站的用户行为数据，包含 7279 个用户对 37,232 个图书项目的 749,401 个评分和 2,056,487 个标签分配活动。两个数据集的评分数值都从 0 到 5，分数步长为 0.5 分。

3.3 框架概述

本章提出的基于异质信息网络的跨领域融合推荐框架 HecRec，通过结合基于 HIN 的推荐算法和跨领域推荐算法，既能有效挖掘目标域异质信息，又能引入辅助域信息增加有效信息量，来更准确的分析目标用户的个性化偏好。该算法能够解决目标域行为数据稀疏对推荐效果的限制，解决冷启动问题，提高评分预测模型的效果。在 HecRec 框架中，有目标推荐领域$\mathcal{D}_t$和辅助域$\mathcal{D}_a$。辅助域作为辅助信息，参与到目标域中，针对用户的进行个性化特征分析，对用户的评分情况进行预测。如图 3-1 所示, 融合框架主要分为三个模块，跨领域异质信息网络构造模块、网络表达学习模块和基于网路表达进行评分预测的模块。首先，两个领域都可以将相关信息表示为单领域 HIN。为了利用来自辅助域的有效信息，我们通过特定的桥梁将两个领域连接起来构成跨领域异质信息网络（Cross-domain HIN）。然后通过设计一系列单领域和跨领域的元路径，挖掘网络中的异质信息，训练节点的嵌入式表达。最后，结合获得的异质信息的表达进行个性化评分预测。在下面章节中会我们会详细介绍各个模块的具体情况。

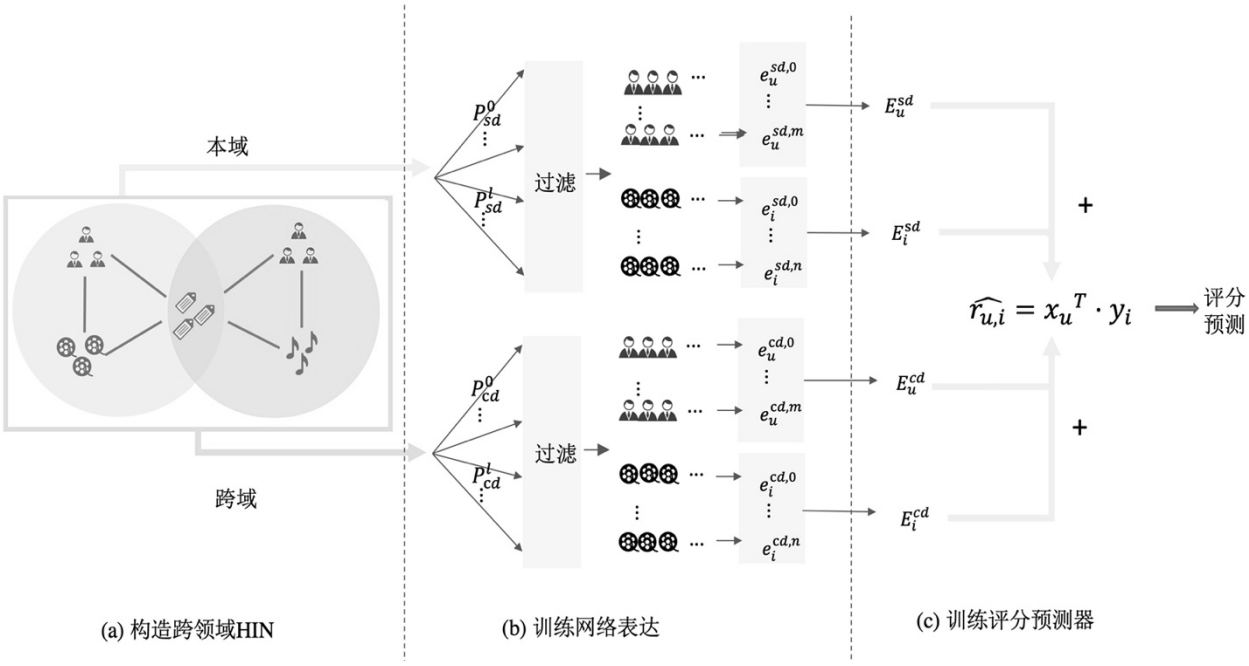

图 3-1 融合框架结构

Figure 3-1 The overview of HecRec

3.4 跨领域 HIN 表达学习

基于异质信息网的推荐框架需要解决两个问题：信息的提取和应用。信息的提取过程需要将网络中的异质信息提取到嵌入式向量中，信息应用过程需要考虑推荐场景，将嵌入式异质信息融合应用于推荐算法，尽可能利用 HIN 中挖掘的有效信息帮助进行用户偏好分析。本节内容主要介绍跨领域 HIN 的表达学习过程，也就是图 3-1 中的前两个模块。

3.4.1 网络构造

跨领域 HIN 表达学习的第一步，就是考虑如何利用不同领域间的信息。如 2.4 跨领域推荐技术中介绍，跨领域推荐主要分为两种信息利用方式，一种是进行特定共享信息的提取，抽象和迁移，如 CBT[5]。另一种是通过整合和利用分布在不同领域中的相关知识进行领域间合作。本文旨在通过挖掘丰富多样的异质信息提高推荐效果，因此采用泛化性能更好的第二种方案，即通过整合来自目标域和辅助域的相关异质信息来挖掘有效信息。

要通过联合相关知识进行领域间合作，需要将目标域和辅助域的数据整合起来。不同领域内部存在多种节点和相关关系，可以表示为单领域内的异质信息网络，我们需要寻找合适的桥梁，将单领域的 HIN 连接为一个完整的跨领域异质信息网络，然后才可以采用基于网络的方式对跨领域异质信息进行提取和利用。所以算法第一阶段的难点就在于，找到稳定的网络间桥梁，保证有效的信息传输。

在本文的工作中，我们采用用户贡献的文本标签（User-contributed Tags）作为

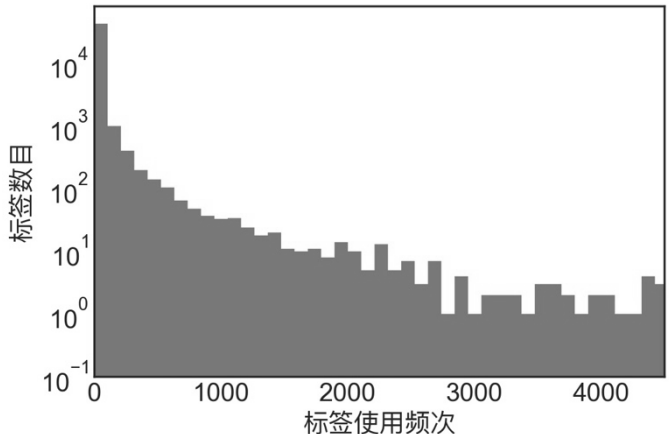

图 3-2 标签使用频率分布情况

Figure 3-2The Frequency Distribution of labels

域间连接桥梁。在一些相关工作中[39, 40, 41]，将领域间重叠的用户和项目作为桥梁，在不同数据集中，通过硬件标识等方式识别出对应用户，或者对同一个数据集进行拆分，来模拟具有相同用户的不同领域。在本文工作中，我们认为重叠用户和项目群对于跨领域的情况并不是广泛存在的，这样的桥梁无法适用于大部分跨领域环境。因此采用网络中的其他节点作为桥梁。通过数据观测，我们发现，有关联价值的领域之间，如本文工作采用的电影和图书数据之间，共享相当比例的标签数据，并且领域中用户高频使用的标签，主要集中于这部分公共标签中。（数据集中标签频率分布和公共标签的占比情况分别如图 3-2 和 3-3 所示。）也就是说，公共标签中包含各数据集中大部分标签数据所包含的信息。而这些标签又能表示不同领域间的用户在偏好上的关系。比如在本文的数据集中，电影和图书之间具有公共的流派、相关人物和目标受众等等，这些共享的信息都可能出现在用户自由创建的标签库中。此外，标签中还包括很多用户的主观信息，这些信息表达了用户对于项目的一些直接感受，而用户的感受往往也是个性化推荐中需要重点关注和分析的内容。

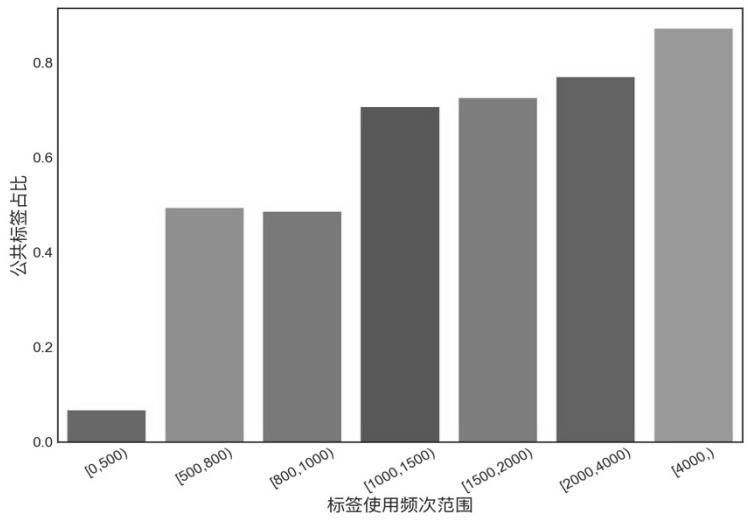

图 3-3 公共标签占比

Figure 3-3 The Proportion of the Common Tags

基于以上假设，我们采用领域间用户贡献的公共文本标签作为桥梁，连接辅助域和目标域的单领域异质信息网络，构造跨领域 HIN。如图 3-4 所示，最左侧分别为电影域和图书域，包含用户、项目和标签三种节点的网络示例，分别用单领域异质信息网络进行建模，其中U_t和I_t表示目标域中节点类型，U_a和I_a表示辅助域中节点类型，两个单领域 HIN 基于公共标签T_2和T_4连接成完整的跨领域异质信息网络。

3.4.2 表达学习

对跨领域异质信息网络的表达学习是本章工作中的重要环节，获得的嵌入式表达需要能够包含不同来源的有效信息。对于跨领域异质信息网络而言，不同于现有的基于异质信息网络的单领域推荐，网络中存在两个单领域网络，子网之间由标签节点作为桥梁进行连接。为使有效信息能够通过桥梁节点进行稳定的迁移并提取嵌入式表达，我们采用元路径方式进行表达学习。我们分别设计了目标域中的元路径和针对跨域信息迁移的元路径，以捕捉跨领域异质信息。另外，我们还发现，在基于元路径的网络表达算法中，承载多种信息的复用节点上存在的信息冲突现象。为了解决这种问题，我们提出“立交桥”的信息整合概念，并据此对异质信息进行整合利用。本节内容主要包括两部分，首先我们设计一系列作用于跨领域异质信息网络的元路径，获得基于元路径的原始节点向量表示；之后介绍对原始节点向量进行“立交桥”式处理获得最终用于推荐的跨领域异质信息网络节点表达。

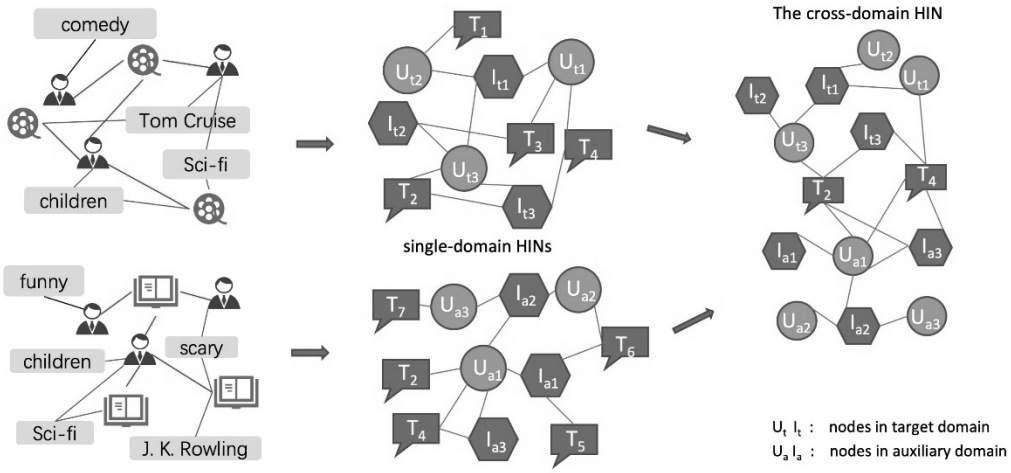

图 3-4 构造跨领域 HIN

Figure 3-4 Constructing the cross-domain HIN

1) 基于元路径的原始向量表达

节点序列获取。为了捕捉跨领域异质信息网络中节点之间的复杂关系，我们在目标域内和域间定义了一系列元路径。与现有的单领域中，基于 HIN 的推荐算法不同，我们设计了八种不同类型的元路径，如表 3-1 所示，其中一半元路径类型用于捕捉域内异质信息，另一半元路径挖掘跨领域的节点关系。作为连接域间进行信息传递的桥梁节点，标签类型节点存在于六种不同的元路径规则当中。

元路径可以帮助我们在异质信息网络中找到结构上没有直接相连，但彼此之间具有相似性的节点关系。以表 3-1 中的元路径p_{cd}^{1}为例，它的路径规则为$U_t - T - U_a - T - U_t$，是跨域使用的元路径。在图 3-5 中所示的网络中，以U_{t3}为开始节点进行随机游走，经过四跳$U_{t3} - T_4 - U_{a1} - T_2 - U_{t4}$到达节点$U_{t4}$。所以$U_{t4}$是$U_{t3}$基于元路径$p_{cd}^{1}$获得的相似用户。此外在图 3-5 中，我们分别基于表 3-1 中用于挖掘相似用户对的四种元路径，即p_{sd}^{0}、p_{sd}^{1}、p_{cd}^{0}和p_{cd}^{1}，生成了四条对应的节点序列，其中两条在目标域中进行随机游走，另外两条通过标签节点引入了域外辅助知识。异质信息网络中基于元路径的随机游走策略在 2.2 节中已有相关介绍，这里不再进行详细说明。

表 3-1 元路径设计

Table 3-1 Meta-path for the cross-domain HIN.

	元路径	路径规则	相关关系
目	p_{sd}^{0}	$U_t - I_t - U_t$	喜欢同一个项目的用户对
标	p_{sd}^{1}	$U_t - T - U_t$	采用同一个标签的用户对
域	p_{sd}^{2}	$I_t - U_t - I_t$	被同一个用户喜欢的项目对
内	p_{sd}^{3}	$I_t - T - I_t$	被同一个标签标记的项目对
跨	p_{cd}^{0}	$U_t - T - I_a - T - U_t$	与共同标记过I_a的两个标签相关的用户对
领	p_{cd}^{1}	$U_t - T - U_a - T - U_t$	与同时被U_a使用过的两个标签相关的用户对
域	p_{cd}^{2}	$I_t - T - I_a - T - I_t$	与同时标记过I_a的两个标签相关的项目对
	p_{cd}^{3}	$I_t - T - U_a - T - I_t$	与同时被U_a使用过的两个标签相关的项目对

由于系统的目标是对用户进行个性化评分的预测，特征分析和表征训练的对象是目标域的用户和项目两种节点类型，即U_t和I_t。因此在基于元路径进行随机游走获得对应的节点序列之后，我们对节点序列进行过滤处理，只保留对应元路径中与第一个节点类型相同的目标类型节点。例如，基于元路径p_{sd}^{0}获得的原始节点序列$U_{t1} - I_{t1} - U_{t3} - I_{t2}\cdots$，过滤掉中间节点后，获得$U_{t1} - U_{t3}\cdots$。

训练原始节点向量。在获得过滤后的节点序列之后，我们同样采用 SkipGram 算法进行嵌入式节点向量的学习，目标函数如式（2-11）所示。另外，由于过滤之后，节点序列变为同质的U_t或I_t序列，序列内相邻的节点间具有元路径代表的相关关系，所以训练过程中窗口长度为 2。

经过训练之后，对于目标域中每个用户或项目节点，获得对应元路径的节点向量。如图 3-6 所示，m和n分别是跨领域 HIN 网络中设计的目标域内和跨域的元路径种类个数，由表 3-1 所示，本文工作中m和n的值均为 4，经过向量化表达训练之

后，用户和项目分别获得本域和跨域的向量集合，其中$e_u^{sd,0}, e_u^{sd,1}$…表示基于用户为起始节点的域内元路径训练的 N 维嵌入式向量的集合，$e_u^{cd,0}, e_u^{cd,1}$…表示基于用户为起始节点的跨域元路径训练的 N 维嵌入式向量的集合。也就是说，基于本文中设计的元路径系列，每个用户节点共获得 4 个 N 维嵌入式向量，其中两个基于元路径p_{sd}^0和p_{sd}^1获得，另外两个基于元路径p_{cd}^0和p_{cd}^1获得。对于项目节点类似。原始嵌入式向量的个数随元路径设计情况的变化而不同。

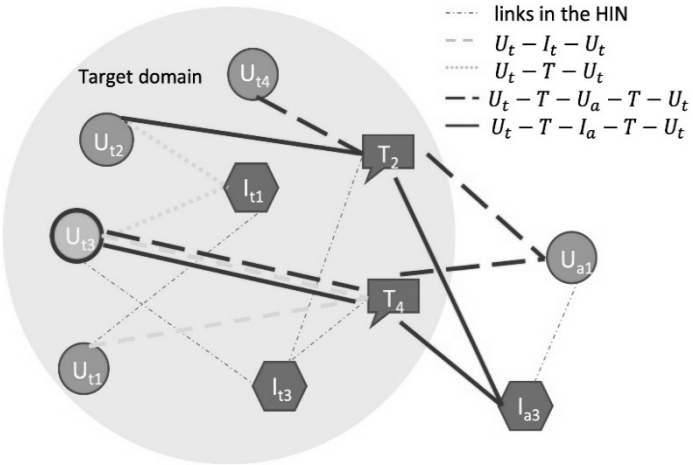

图 3-5 元路径作用示例

Figure 3-5 Explanation of meta-paths.

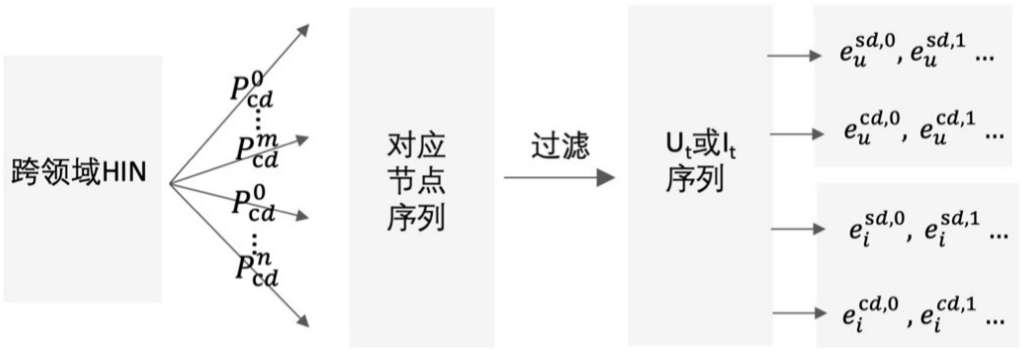

图 3-6 获得原始节点向量流程图

Figure 3-6 The process of obtaining the original embeddings.

2）原始节点向量的处理

为了更好的将异质信息应用于个性化评分预测，需要对获得的原始节点向量集合进行融合处理，生成最终用于评分预测的节点向量。

现有的相关工作中，对于不同的向量化表达的融合往往采取线性加权等方式进行。在 Shi 等人[45]的工作中提出采用线性加权融合，非线性融合和个性化融合等公式，将向量通过计算进行整合，获得融合后统一的向量表示。如式（3-1）所示为个性化线性融合公式，其中$M^i e^i + b^i$表示对向量进行线性变换，w表示个性化加权因子。

$$E = g\left(\{e^i\}\right) = \sum_i w\left(M^i e^i + b^i\right) \tag{3-1}$$

但是在我们的工作中发现，采用这样的融合方式对来源于不同元路径的节点嵌入式表达进行融合，会导致不同来源信息间的冲突。如图 3-5 所示，在图中指出的出的四条节点路径中，节点T_4出现在其中三条路径中。这种程度的复用情况主要是由于在跨领域 HIN 中，标签节点作为桥梁节点，除了承担本域异质信息载体的角色之外，还承担了辅助域信息的迁移任务。所以跨领域节点序列之间，甚至与目标域内的节点序列间，会出现一些重叠的子序列。这些子序列承担的信息传输任务是各不相同的。式（3-1）所示的向量融合计算公式，一定程度上会使特征信息相互抵消，导致各类型元路径无法更好的提供有效信息，甚至会引入无效的噪声影响推荐效果。这种现象在网络规模较小，节点数目缺乏的情况下会更为明显。相关内容将在实验部分进行验证和说明。

为了更好的利用跨领域 HIN 中丰富的异质信息，并同时避免信息冲突和损耗，在进行原始节点向量的处理和整合的过程中，我们引入了“立交桥”式处理的方法。原始向量处理的具体流程为：首先采用函数$f(\cdot)$，引入非线性和个性化因子，对各原始向量进行转化，然后将他们拼接起来。此外，针对来自于目标域和跨域的不同来源的信息做独立并行处理，以最大限度保留节点特征，获得最终的跨领域 HIN 节点向量E_u^{sd}、E_u^{cd}、E_i^{sd}和E_i^{cd}。这种方式下获得的最终的向量表达中，各个元路径像立交桥的各个桥面，分别独立地进行信息的承载和运输。用户节点的向量处理过程可表示为式（3-2）至（3-5）：

$$e_u^{sd,n} = f\left(\{e_u^n\}_{n=i}^{|\mathcal{P}_{sd}|}\right) = \sigma\left(\omega_u^{sd,n}\sigma\left(M^{sd,n}\{e_u^n\}_{n=i}^{|\mathcal{P}_{sd}|} + b^{sd,n}\right)\right) \tag{3-2}$$

$$E_u^{sd} = \left[e_u^{sd,0}, e_u^{sd,1}\cdots\right] \tag{3-3}$$

$$e_u^{cd,n} = f\left(\{e_u^n\}_{n=i}^{|\mathcal{P}_{cd}|}\right) = \sigma\left(\omega_u^{cd,n}\sigma\left(M^{cd,n}\{e_u^n\}_{n=i}^{|\mathcal{P}_{cd}|} + b^{cd,n}\right)\right) \tag{3-4}$$

$$E_u^{cd} = \left[e_u^{cd,0}, e_u^{cd,1}\cdots\right] \tag{3-5}$$

其中$\{e_u^n\}_{n=i}^{|\mathcal{P}_{sd}|}$和$\{e_u^n\}_{n=i}^{|\mathcal{P}_{cd}|}$分别表示用户节点对应本域和跨域的第 i 条元路径获得的原始嵌入式表达向量，$M^{sd} \in \mathbb{R}^{D\times N}$和$M^{cd} \in \mathbb{R}^{D\times N}$分别是本域和跨域的转移矩阵，$b \in \mathbb{R}^D$为偏置向量，$\left(M^{sd,n}\{e_u^n\}_{n=i}^{|\mathcal{P}_{sd}|} + b^{sd,n}\right)$表示对原始向量的线性加权变换，$\omega_u^{sd}$

和ω_u^{cd}是个性化权重矩阵，引入了两层非线性变换 sigmoid 函数$\sigma(\cdot)$，E_u^{sd}和E_u^{cd}中合并的变换后的向量个数分别为$\left|\{e_u^n\}_{n=i}^{|\mathcal{P}_{sd}|}\right|$和$\left|\{e_u^n\}_{n=i}^{|\mathcal{P}_{cd}|}\right|$。

3.5 基于跨领域 HIN 表达的推荐

基于网络表达的推荐算法主要分为网络表达和个性化推荐两个环节。在 3.3 节内容中，我们详细介绍了跨领域 HIN 网络表达算法流程，在获得网络表达之后，如何将嵌入式信息应用于提高个性化评分预测效果的目标上来，就是本节内容的重点。

在上节内容中，我们通过结合跨领域 HIN 网络结构对异质信息进行挖掘和提取，对应目标领域内的用户和项目节点分别获得包含本域和跨域信息的融合后最终的嵌入式向量化表达，即E_u^{sd}、E_u^{cd}、E_i^{sd}和E_i^{cd}。本节中，我们将提出基于矩阵分解的跨领域扩展推荐算法，利用学习到的节点向量中包含的异质信息对用户的个性化评分进行预测。预测评分的计算如式（3-6）所示：

$$\hat{r}_{u,i} = x_u^T \cdot y_i + (E_u^{sd^T} \cdot r_i^{sd} + r_u^{sd^T} \cdot E_i^{sd}) + (E_u^{cd^T} \cdot r_i^{cd} + r_u^{cd^T} \cdot E_i^{cd}) \quad (3\text{-}6)$$

其中，x_u和y_i是用户u和项目i的隐式向量，E_u^{sd}、E_u^{cd}、E_i^{sd}和E_i^{cd}为融合后最终的跨领域异质信息网络的 N 维向量化表达，r_i^{sd}、r_i^{cd}、r_u^{sd}和r_u^{cd}是分别用于和四个网络表达向量配对进行内积的隐式向量，和x_u、y_i相同都需要通过训练获得。引入配对向量的原因是由于，用户和项目的网络表达向量式基于独立的向量空间学习到的，向量彼此之间不同于x_u和y_i，不具有相关关系，可以通过内积模拟用户和项目间的个性化评分行为，不能直接进行内积运算，因此我们用学习到的网络表达向量和配对向量进行内积，引入网络信息，并作为预测评分的一部分引入个性化预测模型。评分预测器主要由三部分组成，如式（3-6）所示，除隐式向量的内积项之外，式中第二项表示目标域中的异质信息的引入，第三项为跨域信息。

我们基于（3-6）进行向量训练，并基于训练结果进行评分预测。训练过程中的目标函数为加L_2正则的预测评分的平方损失函数：

$$loss = \sum_{r_{i,j}} \left(\hat{r}_{i,j} - r_{i,j}\right)^2 + \lambda(\|x_u\|^2 + \|y_i\|^2 + \|R\|^2 + \|\Phi\|^2) \quad (3\text{-}7)$$

其中R表示配对向量的集合，Φ为向量融合过程中转换函数$f(\cdot)$中的相关参数集合，包括 M、b 和个性化权值ω。

3.6 实验验证

本节内容主要介绍实验及评估结果，包括实验的数据集和参数设置、相关工作对比及冷启动对比两部分实验。通过实验，我们主要验证了三方面的问题，即跨领

域辅助信息引入的必要性，“立交桥”式向量融合处理的有效性和融合框架解决冷启动问题的能力。

3.6.1 实验设置

数据集设置。在本章的工作中，选择使用两个数据集的前 5000 个用户和项目的子数据集进行实验。一方面，这样的数据集在跨领域推荐的相关研究里已经是较大规模了，另一方面这样处理为了和对比算法保持一致。数据集的相关统计特征如表 3-2 所示。可以看出电影数据更密集，为了更好的验证算法在冷启动中的表现力，在本章工作中，我们将电影数据集用作辅助域，图书数据集用做目标

表 3-2 数据集的统计特征

Table 3-2 Statistics of the Datasets

	用户数	项目数	评分数	稀疏度	公共标签数
LibraryThing	4974	4998	493460	0.724%	699
MovieLens	4999	4796	581684	2.426%	

域。另外，在需要对用户行为进行二值化处理时，采用 3 分作为阈值，对用户是否喜欢某一项目进行划分。在进行跨领域异质信息挖掘的过程中，我们过滤掉评分不高于三分的用户-项目对，在学习评分预测器的训练过程中，采用数据集完整的评分数据。我们对标签数据进行了处理和过滤，仅保留英文字母和数字内容（有少量低频标签以其他语言字符出现），并且过滤处理后内容长度不超过 1 个字符，或者使用频次不超过 5 次的标签内容。

参数设置。实验过程中，我们每次随机选择 80%的评分数据及相关的异质信息进行训练，20%的数据进行测试，每组实验重复 10 次取平均值作为最终表现。随机游走序列长度取 40，和对照工作[33]一致。网络表达的向量维度和评分预测的隐式向量分别为 128 和 30，正则参数为 0.1。

3.6.2 相关工作对比实验

为了证明本章提出算法的优越性，我们将进行实验对本章提出的 HecRec 框架及以下相关工作的性能进行比较。在这些选定的对比算法中，UBCF，IBCF 和 PMF 被视为传统的推荐方法，ETagiCDCF 是跨域推荐研究领域中较新提出的算法。HERec 是基于异质信息网络的推荐的最新算法，另外我们还引入了它的变体

$HERec_{cd}$ 来进一步验证 HecRec 的优越性。

UBCF：基于用户的协同过滤算法[46]，我们通过多次验证，取最优实验效果的参数，邻居个数设置为 50。

IBCF：基于项目的协同过滤算法[47]，邻居个数大小同样设置为 50。

PMF：概率矩阵分解模型[44]，它在基础 MF 算法上引入概率模型进行进一步优化，对评分矩阵进行低维分解，再基于特征矩阵去预测评分矩阵中的未知值。

ETagiCDCF：较新提出的跨领域推荐框架（Enhanced Tag-induced Cross Domain Collaborative Filtering），该框架是基于 2.4 章节中介绍的 TagCDCF 推荐框架的优化框架。框架核心是基于标签计算跨领域用户和项目的相似度，然后通过跨领域的 PMF 扩展算法进行评分预测。除此之外，ETagiCDCF 通过聚类，不仅利用公共标签内容，同时也利用了各领域特有的标签内容。

HERec：基于异质信息网络的个性化推荐框架（Heterogeneous Information Network Embedding for Recommendation），与本章提出的框架不同，推荐在单一领域进行，没有引入跨域辅助域信息，网络表达向量的处理与本文的“立交桥”式处理不同，采用非线性加权方式直接进行融合。

$HERec_{cd}$：为了更好的验证本章中提出的“立交桥”式处理方案的提高效果，对比实验中提出 HERec 框架的变体框架 $HERec_{cd}$。此框架中按照 HecRec 方法引入辅助域信息，但跨领域异质信息网络表达的融合和处理仍按照 HERec 原方案进行。我们按照 3.4.1 中描述进行参数设置，基于数据集对上述相关工作和本章提出的 HecRec 进行实验，看不同算法框架在 MAE 和 RMSE 指标上的表现，实验结果如表 3-3 所示。

表 3-3 相关工作的对比

Table 3-3 Performance of HecRec and Baselines

	IBCF	UBCF	PMF	ETagiCDCF	HERec	$HERec_{cd}$	HecRec
MAE	0.7248	0.6794	0.6797	0.6789	0.6563	0.6562	0.6384
RMSE	0.8808	1.0099	0.8788	0.8556	0.8481	0.8482	0.8244

由表 3-3 可以看出，HecRec 在 MAE 和 RMSE 上的表现明显优于其他相关工作。和 HecRec 相同，ETagiCDCF 同样是基于标签的跨领域推荐算法，它的表现和 PMF 等传统推荐算法相比也有一定的提升，这说明了辅助域信息在推荐领域的有效性。同时我们注意到，HERec 作为单领域推荐算法，提取的异质信息也仅限于标签信息，但是算法表现和 ETagiCDCF 相比却有显著提升。这说明，采用设计多条元路径，基于随机游走的方式对异质信息网络中的信息挖掘更有效。这是由于，

和多数跨领域推荐和迁移学习相关算法相同，ETagiCDCF 算法迁移的共享知识是单一的，仅基于标签对用户和和项目的相似度进行分析，而元路径的设计更为灵活且全面，因此提升效果更好。将基于元路径的 HIN 网络表达方案引入跨领域推荐场景，能提取出多种相关关系，使跨领域迁移的共享知识更加丰富。值得注意的是，HERec 算法的变体 $HERec_{cd}$，虽然引入了跨领域信息，元路径设计也和 HecRec 相同，但是和 HERec 相比效果却几乎没有提升。这是由于基于元路径获得的原始网络表达的融合方式没能有效的利用提取的异质信息。如 3.3.2 所述，当元路径中节点类型出现复用情况时，基于不同元路径获得的节点向量对应提取了不同的相关特征，采用加权融合的方式给网络表达阶段提取的信息造成了损失，因此虽然同样引入了辅助域信息，$HERec_{cd}$ 依旧无法获得信息带来的收益。

通过对比实验结果，我们验证了 HecRec 框架的优势，并且通过分析各相关工作的表现，验证了跨领域采集信息的必要性、融合框架的优势以及“立交桥”式向量处理的优势。

3.6.3 冷启动对比试验

本文提出的推荐框架 HecRec 主要目的是希望通过采用跨领域和异质信息来改善个性化推荐的冷启动问题，因此本节主要介绍 HecRec 框架在不同活跃度用户群上的表现，验证其冷启动问题的解决能力。

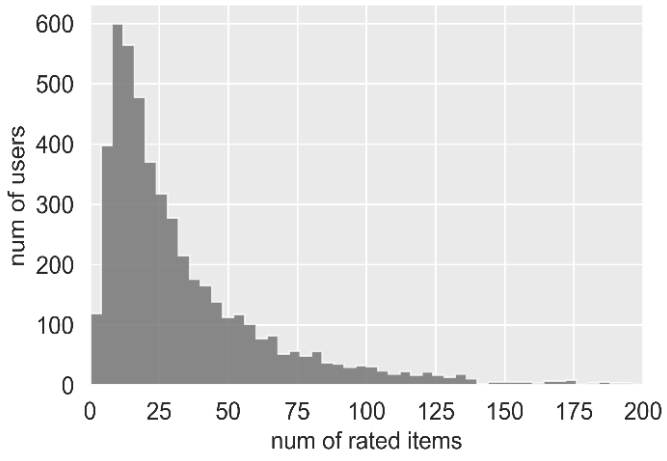

图 3-7 用户评分个数的分布情况

Figure 3-7 The Distribution of Numbers of User’s Rating.

首先对用户基于评分活跃程度进行分组。用户评分量的分布统计情况如图 3-7 所示。基于统计结果，我们将评分数目少于 30 的用户分为三个个群组，每个群组的用户规模如图 3-8 所示。用户评分数目越少，用户活跃度月底，三个用户群组随评分数目的减少，冷启动现象更为明显。我们在评分少于 30 次的用户的三个群组上，分别对相关工作中表现较好的 HERec、$HERec_{cd}$ 与我们的 HecRec 框架进行比较实验，实验结果如表 3-4 所示。从表中实验结果，可以明显看出，相比 HERec 的效果提升，用户活跃度越低，算法的提升效果越明显。在用户评分个数不超过 10 次的群组数据中，HecRec 框架相比较 HERec 提高了近 6.9%。

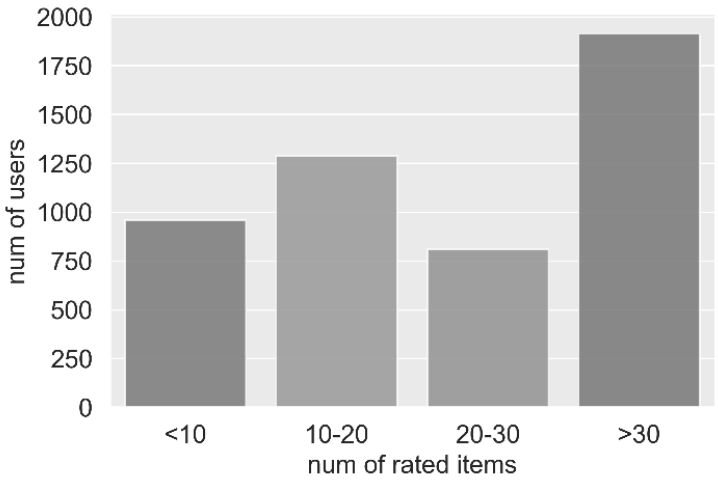

图 3-8 用户群组及其用户量

Figure 3-8 Different User Groups and the Numbers of users

表 3-4 冷启动问题改善能力

Table 3-4 The Ability of Solving Cold-start Problem

用户群	Metirc	HERec	$HERec_{cd}$	HecRec	Improve(%)
(0,10]	MAE	0.9017	0.8930	0.8398	6.8666
	RMSE	1.1640	1.1592	1.0993	5.5625
(10,20]	MAE	0.6899	0.6870	0.6803	1.4050
	RMSE	0.9162	0.9127	0.9077	0.9310
(20,30]	MAE	0.7063	0.7066	0.7033	0.4329
	RMSE	0.9281	0.9280	0.9278	0.0322

3.6.4 实验结论分析

对于本章提出的融合框架的效果评估，我们主要关注以下三个方面：

（1）跨域引入辅助信息的必要性

我们在表 3-2 中可以看到，基于 PMF 的跨领域算法 ETagiCDCF 也在引入了辅助域信息之后推荐效果获得了改善。同样的，引入了辅助域信息的 $HERec_{cd}$ 在较为稀疏的数据集上和 HERec 相比也是有一定提升的。但是由于采用了加权融合的向量处理，因此对辅助信息的挖掘不够充分，与 HecRec 相比，效果有明显的差距。

（2）提出的“立交桥”式融合处理方式的有效性

表 3-3 中，通过对比与 HERec 及其变体的表现情况可以清晰的得出结论。在数据稀疏情况最为明显的数据集，及用户评分次数少于 10 的情况下，目标域对有效性息的缺乏较为明显，对辅助信息的吸收效果受信息处理方式的影响较小，因此在这个子数据集上，原有的处理方法仍然能够有一定程度的提升。但随着稀疏性减弱，信息挖掘的难度和要求提高，辅助数据引入的收益就很不明显了。总和整体数据集来看，表现基本和不引入辅助信息没有差别。相反 HecRec 中辅助信息引入的收益是显著的。所以“立交桥”式融合处理方式在跨领域知识迁移的过程中起到了重要的作用。

（3）HecRec 框架缓解冷启动问题的能力

HecRec 框架中，我们通过结合跨领域和基于异质信息网络两种推荐方式，进行相互补充来解决推荐冷启动问题。从表 3-4 可以看出，HecRec 的提升效果随数据稀疏性变强而更加显著。最稀疏的子数据集上，HecRec 的 RMSE 和 MAE 指标和 HERec 相比，分别提升了 6.6%和 5.5%。

3.7 本章总结

在本章中，我们提出了基于异质信息网络的跨领域融合框架 HecRec，来解决个性化推荐系统中的冷启动问题。具体工作如下：

（1）我们将基于 HIN 的推荐和跨领域推荐方法进行了有机融合，有效的引入跨领域异质信息弥补用户行为数据的不足。基于 HIN 的元路径网络表达方法丰富了跨领域推荐中知识迁移的丰富性，而辅助域数据的引入也增加了可利用的有效异质信息。提高了模型效果。

（2）我们还提出“立交桥”式的网络表达向量处理方法，有效地解决了常用融合方法造成的异质信息挖掘不充分的问题，最大程度地发挥了引入跨领域异质信息的优势。

（3）通过实验可以看出，HecRec 算法与表现最好的相关工作 HERec 相比，H 在 MAE 和 RMSE 上的表现分别提高了 2.7%和 2.8%，在最稀疏的子数据集上，则分别获得了 6.9%和 5.5%的提升。

4 基于网络嵌入传播层的 Top-K 推荐模型

针对个性化推荐中的冷启动问题，本章针对 Top-K 列表推荐场景，提出基于网络嵌入传播层的推荐模型 EPCDRec（Embedding Propagation Layer based cross-domain Recommendation System），对异质信息及跨领域信息进行挖掘。在模型训练的过程中，网络节点的表达学习不仅考虑了包含跨领域异质信息的网络结构特征，同时也结合了用户的个性化偏好信息，实现了端到端结构的推荐模型。下面具体介绍算法细节并通过实验验证模型效果。

4.1 模型概述

端到端模型（end-to-end models）是输入原始数据，输出最终结果的模型。模型从输入到输出需要连续可导，以神经网络的形式实现。而如第三章 HecRec 框架，基于网络结构表达的推荐算法一般较难实现端到端的模型结构。相关工作证明，网络结构特征并不能完全等同于用户个性化特征 [58]，基于 skip-gram 等算法获得的网络节点表达无法直接用于个性化推荐，这类算法往往分模块多阶段进行。比如在第三章的 HecRec 框架中，用户和项目的网络结构的向量化表达是以尽可能缩小相邻节点间的差距为目标进行训练的，而不是以实际的个性化评分为目标。所以E_u和E_v向量只作为中间量，作用在扩展的 MF 算法中进行个性化评分的预测。所以，实现端到端的推荐系统，需要在神经网络结构中学习网络节点的向量化表达，并且学习到的向量表达要同时包含用户的个性化信息，以识别用户偏好为目标。本章中我们提出采用网络嵌入传播层对跨领域异质信息进行挖掘，实现端到端的推荐模型。

网络嵌入传播层（Embedding Propagation Layers，EPL）是由 Wang 等人在文献[49]中提出，用于挖掘用户和项目间多阶连通关系的神经网络单元。用户和项目间有直接互动行为时，为一阶或单阶相关。而用户和项目间还存在非直接的互动行为关系。如用户 U_a 和用户 U_b 喜欢同一个项目 I_c，则 U_b 喜欢的另一个项目 I_b 和 U_a 之间就为二阶连通关系。每层传播层对用户和直接相关的项目间信息传递的过程进行模拟。多层传播层的复合堆叠则实现了对网络中多阶连通关系的信息挖掘。与现有的基于网络表达的推荐算法不同，网络嵌入传播层的使用将网络拓扑结构以神经网络的形式表现出来，挖掘网络拓扑信息。

本章工作中，我们在 Wang 等人研究[48]的基础上，提出基于网络嵌入传播层（Embedding Propagation Layers，EPL）的推荐模型 EPCDRec，用于 Top-K 列表推荐。我们采用基于 EPL 的神经网络结构，挖掘跨领域 HIN 的拓扑信息，并通过贝

叶斯个性化排序（BPR）算法进行嵌入式向量的训练，使最终获得的网络节点表达既包含跨领域 HIN 的网络结构特征，又适用于对用户的个性化推荐。

4.2 网络结构嵌入传播层

本章工作中，我们通过网络嵌入传播层来模拟跨领域 HIN 中的信息传播过程，下面首先对网络嵌入传播层进行介绍。

单阶嵌入传播层。单层的嵌入传播层用于模拟一阶邻接关系，即直接相关的节点间关系。对于只包含用户和项目两种节点的网络，任意一个节点包含的有效信息，包括由邻居节点传递的信息以及节点自身包含的信息。因此，受 Wang 等人[49]工作的启发，我们定义用户节点 u 的一阶向量为：

$$e_u^1 = g(m_{uu} + \sum_{i \in \mathcal{N}^u} m_{iu}) \tag{4-1}$$

其中m_{uu}表示节点自身包含的信息，m_{iu}表示节点 u 的其中一个邻居节点的邻接信息，激活函数$g(\cdot)$采用常用的 LeakyReLU 函数。其中m_{iu}定义为：

$$m_{iu} = f(e_u, e_i) = \frac{1}{|\mathcal{N}^u||\mathcal{N}^i|}\left(W^1 e_i + W^2(e_u \odot e_i)\right) \tag{4-2}$$

从节点 i 传递给节点 u 的信息m_{iu}包括两部分，节点间关系$e_u \odot e_i$，和节点 i 自身信息，其中，e_u和e_i表示当前节点的嵌入式表达，$\odot$表示向量的逐元素乘积运算，W^1和W^2两个参数矩阵作为权重选择因子对有效信息进行筛选，$1/|\mathcal{N}^u||\mathcal{N}^i|$是连边的衰减因子。式（4-1）中的$m_{uu} = W^1 e_u$，和 i 共享权重矩阵W^1。和用户节点的一阶嵌入式输出e_u^1类似，e_i^1可通过对应的计算获得。

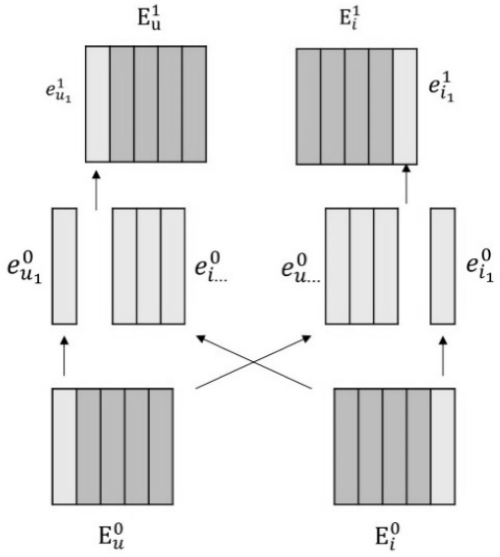

图 4-1 单阶嵌入传播层示意图

Figure 4-1 Illustration of first-order Propagation layer.

多阶嵌入传播层。单阶的嵌入传播层提取了节点间的直接关系特征。通过 n 层嵌入传播层的叠加，可以获取到节点经过 n 跳之后的节点关系。第 n 层传播层的用户节点的嵌入式向量为：

$$e_u^n = g(m_{uu}^n + \sum_{i \in \mathcal{N}^u} m_{iu}^n) \tag{4-3}$$

其中，

$$m_{iu}^n = f(e_u^{n-1}, e_i^{n-1}) = \frac{1}{|\mathcal{N}^u||\mathcal{N}^i|}\left(W^1 e_i^{n-1} + W^2(e_u^{n-1} \odot e_i^{n-1})\right) \tag{4-4}$$

$$m_{uu}^n = W^{1n} e_u^{n-1} \tag{4-5}$$

单阶和高阶的嵌入传播层的结构原理如图 4-1 和 4-2 所示。图 4-2 以节点 u 和 i 为例，描述了经过两层嵌入传播层构成的神经网络结构后，将每层输出的表达进行拼接处理，分别获得用户和项目最终的嵌入式表达结果。

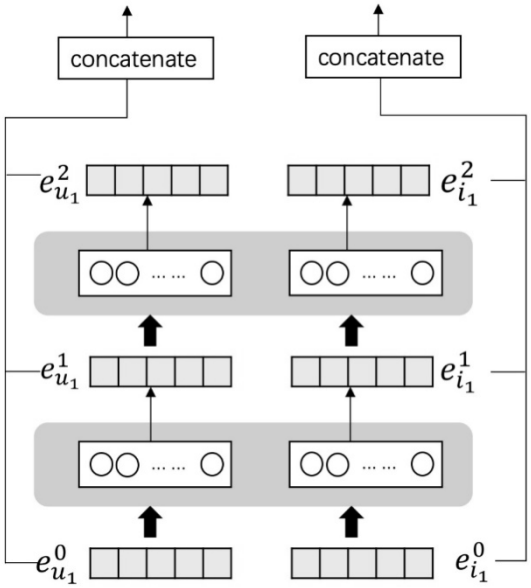

图 4-2 两阶嵌入传播层示意图

Figure 4-2 Illustration of the neural network of two-order Propagation layers.

4.3 基于跨领域异质信息的 Top-K 推荐

上节内容主要介绍了在用户和项目两类节点的网络中，如何用基于网络嵌入传播层的神经网络结构对网络中的信息传递过程进行建模。而本文主要的研究目标是对跨领域异质信息网络的挖掘。所以本节工作中，我们首先对跨领域异质信息网络进行处理，然后基于网络嵌入传播层，构造用于挖掘跨领域 HIN 的神经网络

结构，并基于输出的嵌入式节点表达进行对用户的个性化 TopK 列表推荐。

4.3.1 模型结构

本章工作中我们同样采用标签节点作为桥梁对目标域和辅助域信息进行处理，来构造跨领域异质信息网络，如图 3-4 所示。网络中的有效信息分三部分进行提取，分别是目标域中用户和项目间的直接互动行为信息，目标域中的异质信息和跨领域信息。其中用户和项目间互动行为信息可以直接由多层网络嵌入传播层处理。异质信息和跨领域信息则需要先通过元路径的方式进行预处理。

为了利用嵌入传播层的结构，我们将异质信息和跨领域信息处理为用户和项目的直接关系。如图 4-3 所示，对于异质信息，我们通过元路径$U_t - T - I_t$获取包含异质信息的用户项目对，即和相同标签具有连边关系，但彼此之间无连边的用户和项目。然后将获得的的用户项目关系，输入独立的基于 EPL 的神经网络结构中进行训练。对于跨领域信息，则使用元路径$U_t - T - I_a - T - I_t$和$U_t - T - U_a - T - I_t$来提取相关的用户项目对，并同样构造独立的网络嵌入传播层网络结构。

在分别获得针对不同信息提取出用户项目对之后，我们将三类信息分别送入各自独立的传播层网络进行训练，最大化保留不同特征信息的完整性。直接行为信息、异质信息和跨领域信息对应的网络结构中的层数设置通过具体的实验结果确定。最后，对获得的用户和项目的嵌入式表达 E_U 和 E_I 做内积处理获得排序评分。

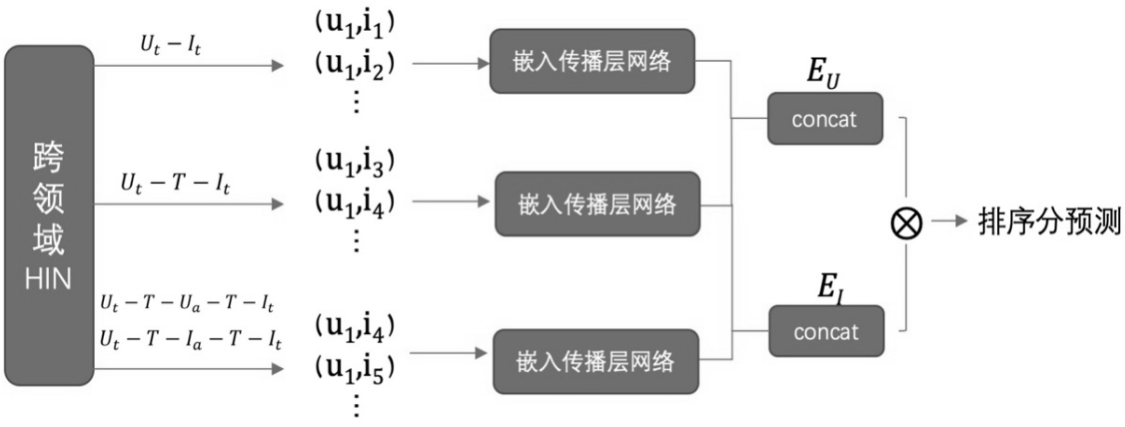

图 4-3 模型结构

Figure 4-3 Illustration of the model.

4.3.2 模型训练

模型训练部分采用 BPR 算法。需要明确的是，不同于第三章的 HecRec 框架，EPCDRec 的排序评分不是对用户实际评分的预测结果，而是用于筛选 TopK 推荐列表的指标。我们采用 BPR 算法对模型进行训练，将用户和项目间的负向关系通过目标函数引入模型训练的过程当中。在这个过程中，节点嵌入式表达的学习以个性化推荐为目标，在通过嵌入传播层进行网络结构特征挖掘的同时，也提取了用户的个性化偏好信息。模型的目标函数为：

$$Loss = \sum_{(u,i,j)} -ln\sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right) + \lambda\|\Theta\|_2^2 \tag{4-6}$$

其中，$\hat{y}_{ui}$和$\hat{y}_{uj}$分别是用户 u 对正向反馈项目和抽样的负向反馈项目的预测排序分值，Θ是相关训练参数，包括初始化嵌入式向量E_U^0和E_I^0，以及权重矩阵W^1和W^2。

4.4 实验验证

本章实验使用的数据集与第三章相同。为了模拟用户评分缺失的情况，我们将评分大于 3 的用户项目关系划定为用户正向行为关系，小于 3 为负向关系。每次随机选择 80%的评分数据及相关的异质信息进行训练，20%的数据进行测试，每组实验重复 10 次取平均值作为最终表现。参数方面，网络结构的嵌入传播层的神经元个数为 16，层数由后续实验结果确定，嵌入式表达维度为 64，正则项系数λ取 1e-5，学习率为 0.0001。

4.4.1 网络层数设置实验

模型中多阶嵌入传输层构成的神经网络结构用来挖掘用户项目二元组中包含的网络结构信息。我们通过对传播层的逐次叠加实验对每部分的具体传播层层数进行确定。实验中我们将 TopK 中 K 取 20。对于$U_t - I_t$关系对，即用户直接行为关系，网络层叠加到 3 层时，模型表现就开始下降，所以取 2 层。然后继续叠加目标领域异质信息，根据$U_t - \mathrm{T} - I_t$获取信息对应的神经网络传输层。最后确定跨领域信息的传输层的层数。

根据表 4-1 所示实验结果，三个传输层网络结构的层数分别设置为 2，1 和 1。值得注意的是，异质信息和跨领域对应的传输层多层叠加后效果没有提升，这是由于这两部分的 U-I 关系本身就是基于元路径获得，因此两层及以上的传输层对应的 U-I 关系对的相关性很弱，所以基于元路径提取的异质信息和跨领域信息只采用了单层嵌入传播层。从表 4-1 中也可以验证，异质信息和跨领域信息的引入都提

高了模型的推荐效果。

表 4-1 嵌入传播层网络的层数选择

Table 4-1 The selection of numbers of layers

叠加$U_t - I_t$层			
	Recall	Precison	Ndcg
+1 layer	0.15631	0.12382	0.21451
+2 layer	**0.15807**	**0.12457**	**0.21685**
+3 layer	0.15649	0.1233	0.21387
$U_t - I_t$层数取 2，叠加$U_t - T - I_t$层			
	Recall	Precison	Ndcg
+1 layer	**0.15995**	**0.12608**	**0.21834**
+2 layer	0.15926	0.12421	0.21895
$U_t - T - I_t$层数取 1，叠加跨领域信息传输层			
	Recall	Precison	Ndcg
+1 layer	**0.16257**	**0.12602**	**0.22345**
+2 layer	0.15283	0.11950	0.21255

表 4-2 相关工作对比

Table 4-2 Performance of EPCDRec and Baselines

Metrics		NeuMF	NGCF	EPRec	EPCDRec
Recall	Top@20	0.09932	0.15807	0.15995	**0.16257**
	Top@50	0.18751	0.31318	0.31558	**0.31668**
	Top@80	0.25487	0.41024	0.41351	**0.41476**
	Top@100	0.29091	0.45786	0.46101	**0.46295**
Precision	Top@20	0.08582	0.12457	0.12608	**0.12612**
	Top@50	0.07008	0.10670	0.10720	**0.10778**
	Top@80	0.06051	0.09327	0.09400	**0.09429**
	Top@100	0.05731	0.08637	0.08720	**0.08723**
Ndcg	Top@20	0.18306	0.21685	0.21834	**0.22345**
	Top@50	0.27879	0.34061	0.34139	**0.34670**
	Top@80	0.33964	0.41609	0.41708	**0.42263**
	Top@100	0.37594	0.45362	0.45513	**0.46028**

4.4.2 相关工作对比实验

为证明模型的优越性，本节实验把 EPCDRec 和相关工作进行对比。我们选择的相关工作中，NeuMF[50]算法是矩阵分解和多层感知机的一个融合推荐算法，矩

阵分解适合抓取乘法关系，多层感知机在学习匹配关系时，通过神经网络的结构训练用户和项目间的非线性关系，但是只考虑用户和项目的直接互动行为信息。NGCF 算法提出了嵌入传播层构成的神经网络结构，还引入了 node dropout 和 message dropout 等措施防止过拟合问题，但是 NGCF 为单领域推荐，没有考虑异质信息和跨领域信息的引入。EPRec 和 EPCDRec 相比，只引入异质信息，不引入辅助域数据。实验过程中，相关工作的神经网络层数和 EPCDRec 保持一致，均为两层，其他参数不变。实验结果如表 4-2 所示。

4.5 本章总结

本章介绍了针对 Top-K 列表推荐场景下提出的基于网络嵌入传播层的端到端跨领域异质信息融合推荐模型 EPCDRec。该模型通过嵌入传播层来处理用户项目间直接关系和基于元路径引入的间接关系，从而综合目标域用户行为信息、异质信息和跨领域信息进行个性化推荐。并通过网络嵌入传播层构成的神经网络结构和 BPR 算法的结合，实现端到端推荐模型。通过我们在 MovieLens 和 LibraryThing 数据集上的实验结果显示：EPCDRec 模型引入异质信息和跨领域信息后，与相关工作 NeuMF 和 NGCF 相比，推荐效果从各个指标上均有明显提升。和 NGCF 相比，K 取 20 的情况下召回率、精准率和 NDCG 分别提升了 2.8%、1.2%和 3.0%。

5 结论

5.1 本文工作总结

为了解决个性化推荐中用户行为数据缺失导致的冷启动下推荐性能下降难题，本文对结合异质信息网络和跨领域的融合推荐算法进行了研究。融合算法的实现主要有以下难点：1）探索信息间有效的融合策略。2）缺乏用于挖掘跨领域异质信息的有效网络表达算法。3）基于异质信息的推荐算法中，端到端模型实现具有一定的难度。为此，我们采用用户提供的标签作为桥梁将目标域和辅助域信息进行融合，分别针对个性化推荐领域的两大应用场景——评分预测和 Top-K 列表推荐，提出了基于异质信息网络表达的跨领域推荐框架 HecRec 和基于网络嵌入传播层的 Top-K 推荐模型 EPCDRec，挖掘有效的跨领域异质信息。

本文的主要工作及贡献如下。

(1) 针对个性化推荐中用户行为数据缺失的问题，提出引入异质信息和跨领域信息作为辅助信息来提高推荐效果。采用标签作为桥梁，构造跨领域异质信息网络，将基于异质信息网络的推荐和跨领域推荐进行有机结合，实现推荐算法的融合。

(2) 针对评分预测推荐场景，提出基于网络表达的融合算法框架以挖掘复杂多样的跨领域异质信息，同时采用了“立交桥式”向量处理方法，避免了信息间冲突造成知识的负迁移情况，以最大化融合框架的优势提高推荐效果。最终融合算法框架的绝对平均误差为 0.6384，比相关工作减少了 2.7%。

(3) 针对 Top-K 列表推荐场景，提出基于嵌入传播层的融合算法模型进行跨领域异质信息的挖掘，并在端到端模型中实现推荐效果的提升。融合模型在真实数据集上的准确率、召回率和归一化折损累计增益分别达到了 0.13、0.16 和 0.22，和相关工作相比分别提升了 2.8%、1.2%和 3.0%。

5.2 未来工作展望

本文的工作为解决个性化推荐领域的用户行为数据稀疏造成的冷启动问题提出了一种新的研究方向，即通过将基于异质信息网络的推荐和跨领域推荐结合起来的方法，改善用户行为数据稀疏对个性化偏好分析的限制。未来的工作希望在能在以下两方面有所突破：一方面是，是需要探索更为有效且稳定的域间桥梁对有效信息进行迁移。另一方面，本文提出的算法中，用户和项目间关系仍采用广泛被应用的内积运算进行模拟，将不同方式获得嵌入式表达直接进行内积作为匹配的量

化结果，未来计划尝试其他形式的互动模型对用户项目关系进行研究。

参考文献

[1] F. Ricci, L. Rokach and B. Shapira, “Introduction to Recommender Systems Handbook”, in Recommender Systems Handbook 1st ed. USA: Springer, 2011, ch. 1, pp. 1-35

[2] Koren Y , Bell R , Volinsky C . Matrix Factorization Techniques for Recommender Systems[J]. Computer, 2009, 42(8):30-37.

[3] M. Sun, F. Li, J. Lee, K. Zhou, G. Lebanon, and H. Zha, “Learning multiple-question decision trees for cold-start recommendation”, in Proc. of WSDM ’13, ACM, Rome, Italy, 2013, pp. 445-454.

[4] Sun Y, Han J. Mining heterogeneous information networks: a structural analysis approach[J]. Acm Sigkdd Explorations Newsletter, 2013, 14(2): 20-28.

[5] Ou M, Cui P, Wang F, et al. Comparing apples to oranges: a scalable solution with heterogeneous hashing[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 230-238.

[6] Shi Y, Zhao X, Wang J, et al. Adaptive diversification of recommendation results via latent factor portfolio[C]//Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 2012: 175-184.

[7] 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.

[8] Y. Desrosiers and G. Karypis, “A Comprehensive Survey of Neighborhood-based Recommendation Methods”, in Recommender Systems Handbook 1st ed. USA: Springer, 2011, ch. 4, pp. 107-144

[9] Y. Koren and R. Bell, “Advances in Collaborative Filtering”, in Recommender Systems Handbook 1st ed. USA: Springer, 2011, ch. 5, pp. 145-186

[10] 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.

[11] Lu Z, Zhong E, Zhao L, et al. Selective transfer learning for cross domain recommendation[C]//Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2013: 641-649.

[12] Moreno O, Shapira B, Rokach L, et al. Talmud: transfer learning for multiple domains[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 425-434.

[13] Ma H, Zhou D, Liu C, et al. Recommender systems with social regularization[C]//Proceedings of the fourth ACM international conference on Web search and data mining. 2011: 287-296.

[14] Li B, Yang Q, Xue X. Transfer learning for collaborative filtering via a rating-matrix generative model[C]//Proceedings of the 26th annual international conference on machine learning. 2009: 617-624.

[15] Ling G, Lyu M R, King I. Ratings meet reviews, a combined approach to recommend[C]//Proceedings of the 8th ACM Conference on Recommender systems. 2014: 105-112.

[16] Rafailidis D, Axenopoulos A, Etzold J, et al. Content-based tag propagation and tensor factorization for personalized item recommendation based on social tagging[J]. ACM Transactions on Interactive Intelligent Systems (TiiS), 2014, 3(4): 1-27.

[17] Hong L, Doumith A S, Davison B D. Co-factorization machines: modeling user interests and predicting individual decisions in twitter[C]//Proceedings of the sixth ACM international conference on Web search and data mining. 2013: 557-566.

[18] 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.

[19] Wang D, Cui P, Zhu W. Structural deep network embedding[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016: 1225-1234.

[20] Pan S, Wu J, Zhu X, et al. Tri-party deep network representation[J]. Network, 2016, 11(9): 12.

[21] Yang C, Liu Z, Zhao D, et al. Network representation learning with rich text information[C]//Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

[22] Zhang D, Yin J, Zhu X, et al. Homophily, structure, and content augmented network representation learning[C]//2016 IEEE 16th international conference on data mining (ICDM). IEEE, 2016: 609-618.

[23] Feng W, Wang J. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012: 1276-1284.

[24] Yu X, Ren X, Gu Q, et al. Collaborative filtering with entity similarity regularization in heterogeneous information networks[J]. IJCAI HINA, 2013, 27.

[25] Yu X, Ren X, Sun Y, et al. Personalized entity recommendation: A heterogeneous information network approach[C]//Proceedings of the 7th ACM international conference on Web search and data mining. 2014: 283-292.

[26] Luo C, Pang W, Wang Z, et al. Hete-cf: Social-based collaborative filtering recommendation using heterogeneous relations[C]//2014 IEEE International Conference on Data Mining. IEEE, 2014: 917-922.

[27] Shi C, Zhang Z, Luo P, et al. Semantic path based personalized recommendation on weighted heterogeneous information networks[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015: 453-462.

[28] Shi C, Liu J, Zhuang F, et al. Integrating heterogeneous information via flexible regularization framework for recommendation[J]. Knowledge and Information Systems, 2016, 49(3): 835-859.

[29] Zheng J, Liu J, Shi C, et al. Recommendation in heterogeneous information network via dual similarity regularization[J]. International Journal of Data Science and Analytics, 2017, 3(1): 35-48.

[30] Shi Y, Larson M, Hanjalic A. Tags as bridges between domains: Improving recommendation with tag-induced cross-domain collaborative filtering[C]//International Conference on User Modeling, Adaptation, and Personalization. Springer, Berlin, Heidelberg, 2011: 305-316.

[31] Zheng J, Liu J, Shi C, et al. Dual similarity regularization for recommendation[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2016: 542-554.

[32] Li B, Yang Q, Xue X. Can movies and books collaborate? cross-domain collaborative

filtering for sparsity reduction[C]//Twenty-First international joint conference on artificial intelligence. 2009.

[33] Hao P, Zhang G, Lu J. Enhancing cross domain recommendation with domain dependent tags[C]//2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2016: 1266-1273.

[34] 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.

[35] Xu L, Wei X, Cao J, et al. Embedding of Embedding (EOE) Joint Embedding for Coupled Heterogeneous Networks[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017: 741-749.

[36] Dong Y, Chawla N V, Swami A. metapath2vec: Scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017: 135-144.

[37] Fu T, Lee W C, Lei Z. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. 2017: 1797-1806.

[38] Chen L, Zheng J, Gao M, et al. TLRec: transfer learning for cross-domain recommendation[C]//2017 IEEE International Conference on Big Knowledge (ICBK). IEEE, 2017: 167-172.

[39] Berkovsky S, Kuflik T, Ricci F. Mediation of user models for enhanced personalization in recommender systems[J]. User Modeling and User-Adapted Interaction, 2008, 18(3): 245-286.

[40] Winoto P, Tang T. If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? a study of cross-domain recommendations[J]. New Generation Computing, 2008, 26(3): 209-225.

[41] Zhang Y, Cao B, Yeung D Y. Multi-domain collaborative filtering[J]. arXiv preprint arXiv:1203.3535, 2012: 725–732.

[42] Loizou, A.: How to recommend music to film buffs: enabling the provision of recommendations from multiple domains. PhD dissertation, University of Southampton (2009).

[43] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix t-factorizations for clustering[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006: 126-135.

[44] Mnih A, Salakhutdinov R R. Probabilistic matrix factorization[C]//Advances in neural information processing systems. 2008: 1257-1264.

[45] Shi C, Hu B, Zhao W X, et al. Heterogeneous information network embedding for recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(2): 357-370.

[46] Herlocker J L, Konstan J A, Borchers A, et al. An algorithmic framework for performing collaborative filtering[C]//ACM SIGIR Forum. New York, NY, USA: ACM, 2017, 51(2): 227-234.

[47] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web. 2001: 285-295.

[48] Wang X, He X, Wang M, et al. Neural graph collaborative filtering[C]//Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval.

2019: 165-174.

[49] 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.

[50] Yu J, Gao M, Li J, et al. Adaptive implicit friends identification over heterogeneous network for social recommendation[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 357-366.

[51] 李娴,赵霞,张泽华,张晨威.基于异质信息网络的模糊推荐算法[J].计算机工程与科学,2020,42(02):334-340.

[52]陶鸿,吴国栋,孙成,查志康,陈海涵.跨领域推荐研究进展[J].长春师范大学学报,2019,38(12):44-54.

作者简历及攻读硕士学位期间取得的研究成果

一、作者简历

尹姜谊，女，1996 年 1 月生。2013 年 9 月至 2017 年 6 月就读于北京交通大学电子与信息工程学院通信工程专业，取得工学学士学位。2017 年 9 月至 2020 年 6 月就读于北京交通大学电子与信息工程学院通信与信息系统专业，研究方向是信息网络，取得工学硕士学位。攻读硕士学位期间，主要从事个性化推荐方面的工作。

二、发表论文

[1] J. Yin, Y. Guo and Y. Chen, "Heterogenous Information Network Embedding Based Cross-Domain Recommendation System," 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 2019.

三、参与科研项目

[1] 活动型社会网络的多重推荐算法研究

[2] 基于异质信息网络的跨领域推荐

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：尹善谊 签字日期：2020 年 6 月 1 日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
个性化推荐；冷启动；异质信息网络；跨领域推荐	公开			
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工学	硕士
论文题名*		并列题名		论文语种*
基于异质信息网络的跨领域推荐系统				中文
作者姓名*	尹姜谊		学号*	17120156
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
通信于信息系统		信息网络	3	2020
论文提交日期*	2020.05.6			
导师姓名*	郭宇春		职称*	教授
评阅人	答辩委员会主席*		答辩委员会成员	
电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） 推荐格式：application/msword；application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数*	51 页			
共 33 项，其中带*为必填数据，为 21 项。				