

**本 科 毕 业 设 计（论文）**



**题目: 异质信息网络中的相似性推荐算法研究**

**姓 名 王睿嘉**

**学 院 计算机学院**

**专 业 计算机科学与技术**

**班 级 2015211307**

**学 号 2015211906**

**班内序号 31**

**指导教师 石川**

**2019 年 5 月**

**北 京 邮 电 大 学**

**本科毕业设计（论文）任务书**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 学院 | 计算机学院 | | 专业 | 计算机科学与技术 | | 班级 | 2015211307 |
| 学生姓名 | 王睿嘉 | | 学号 | 2015211906 | | 班内序号 | 31 |
| 指导教师姓名 | 石川 | | 所在单位 | 计算机学院 | | 职称 | 教授 |
| 设计(论文)  题目 | （中文）异质信息网络中的相似性推荐算法研究 | | | | | | |
| （英文）Research on Similarity Recommendation Algorithm in Heterogeneous Information Network | | | | | | |
| 题目分类 | 工程实践类□ 研究设计类🗹 理论分析类□ | | | | | | |
| 题目来源 | 题目是否来源于科研项目 是🗹 否□ | | | | | | |
| 科研项目名称：基于半结构化/非结构化信息的领域知识网络自动构建技术  合作项目 | | | | | | |
| 科研项目负责人：石川 | | | | | | |
| 主要任务及目标：   1. 阅读异质信息网络综述，了解异质信息网络中的关键概念及相似性推荐算法的研究现状。 2. 进一步调研相关论文，了解现有的高效实现方案。 3. 对比分析现有的相似性推荐算法，选择合适的算法并改进。 4. 根据改进后的算法，设计方案模型。 5. 完成编码工作，正确实现模型并进行实验验证及结果评估。 6. 基于相关文档及实验结果，完成毕业设计论文。 | | | | | | | |
| 主要内容：   1. 调研相似性推荐算法，完成开题工作。 2. 基于对现有算法的分析，选择合适的算法并改进，进一步完成方案的详细设计。 3. 基于经典数据集对方案进行实验验证，并对结果进行分析评估，完成相应中期报告。 4. 完成并提交毕业设计论文、相关文档及源程序。 | | | | | | | |
| 主要参考文献：   1. Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, Philip S. Yu. A survey on Heterogeneous Information Network Analysis. IEEE Transactions on Knowledge and Data Engineering, 2017. 2. Chuan Shi，Chong Zhou, Xiangnan Kong, Philip S. Yu, Gang Liu, Bai Wang. HeteRecom: A Semantic-based Recommendation System in Heterogeneous Networks. KDD 2012. 3. Binbin Hu, Chuan Shi, Wayne Xin Zhao, Tianchi Yang. Local and Global Information Fusion for Top-N Recommendation in Heterogeneous Information Network. CIKM 2018. 4. Xiaotian Han, Chuan Shi, Lei Zheng, Philip S. Yu, Jianxin Li, Yuanfu Lu. Representation Learning with Depth and Breadth for Recommendation using Multi-view Data. APWeb-WAIM 2018. 5. Chuan Shi, Zhiqiang Zhang, Yugang Ji, Weipeng Wang, Philiph S. Yu, Zhiping Shi. SemRec: A Personalized Semantic Recommendation Method based on Weighted Heterogeneous Information Networks. WWW 2018. 6. Binbin Hu, Chuan Shi, Wayne Xin Zhao, Philip S. Yu. Leveraging Meta-path based Context for Top-N Recommendation with A Neural Co-Attention Model. KDD 2018. 7. Chuan Shi, Zhiqiang Zhang, Ping Luo, Philip S. Yu, Yading Yue, Bin Wu. Semantic Path based Personalized Recommendation on Weighted Heterogeneous Information Networks. CIKM 2015. | | | | | | | |
| 进度安排：  第一阶段（01-02周）：阅读相关论文，了解相似性推荐的实现算法，并提交开题报告；  第二阶段（03-04周）：根据调研结果，对比分析现有的相似性推荐算法，选择合适的算法，并在此基础上改进，进一步设计详细的实现方案；  第三阶段（05-08周）：熟悉数据特点及相应算法框架，完成方案算法的编程实现，进行中期检查；  第四阶段（09-11周）：完成模型实验验证模块的编程实现，并对结果进行评估；  第五阶段（12-15周）：完成毕业设计论文的撰写，进行答辩。 | | | | | | | |
| 指导教师签字 | |  | | 日期 | 2019 年 3 月 10 日 | | |

**北 京 邮 电 大 学**

**本科毕业设计（论文）答辩决议书**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学院 | 计算机学院 | | | 专业 | | 计算机科学与技术 | | 班级 | | 2015211307 |
| 学生姓名 | 王睿嘉 | | | 学号 | | 2015211906 | | 班内序号 | | 31 |
| 毕业设计  （论文）题目 | （中文）异质信息网络中的相似性推荐算法研究 | | | | | | | 百分制成绩 | |  |
| （英文）Research on Similarity Recommendation Algorithm in Heterogeneous Information Network | | | | | | | 五级分制成绩 | |  |
| 指导教师姓名 | 石川 | 所在单位 | | | | 计算机学院 | | 职称 | | 教授 |
| 指导教师评语： | | | | | | | | | | |
| 指导教师评分(满分40分) |  | | 签字 | |  | | 日期 | | 年 月 日 | |
| 答辩小组评语：  答辩小组评分： 组长职称： 签字：  (满分60分) 成员职称： 签字：  成员职称： 签字：  成员职称： 签字：  成员职称： 签字：  年 月 日 | | | | | | | | | | |

**北 京 邮 电 大 学**

**本科毕业设计（论文）诚信声明**

本人声明所呈交的毕业设计（论文），题目《异质信息网络中的相似性推荐算法研究》是本人在指导教师的指导下，独立进行研究工作所取得的成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，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邮电大学或其他教育机构的学位或证书而使用过的材料。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

本人签名： 日期：

**异质信息网络中的相似性推荐算法研究**

**摘 要**

推荐是解决信息过载的有效方法，被广泛应用于电子商务和互联网服务中。作为重要的数据挖掘任务，推荐系统包括许多对象类型及对象间的丰富交互关系，可以自然而然地利用异质信息网络（Heterogeneous Information Network）建模。而异质信息网络优越的信息综合能力及丰富的语义信息，进一步使得更准确的推荐成为可能。但是，现有异质信息网络的推荐算法大多聚焦于用户和物品间的交互信息，即购买历史，而这在实际应用中仅能反映用户口味和物品特性的一个方面。因此，为深入挖掘不同方面的语义关系相似性，Neural network based Aspect-level Collaborative Filtering model (NeuACF)设计了基于元路径的推荐算法，使得推荐模型融合多方面语义信息，得到更加全面精确的结果。

本文首先介绍选题背景及相关技术，从宏观角度回顾推荐领域的经典算法。其次，深入分析NeuACF的具体模型结构，并针对其存在的两点问题，提出相应改进方案。具体地，用户和物品的交互可能性利用点积度量，可能导致相似性的传递特征被破坏。为此，本文引入metric learning，利用距离定义的metric捕捉数据间关系，实现更全方位的信息传递。与此同时，NeuACF使用point-wise损失函数，关注于评分信息的绝对数值。但对于Top-N推荐而言，相对的交互可能性更为重要。因此，本文考虑使用基于负采样的pair-wise损失函数，产生排序更合理的推荐列表。最后，本文在推荐数据集Amazon和MovieLens上，与经典算法进行了大量对比实验，验证了所提模型的优越性。

**关键词** 推荐 异质信息网络 度量学习

**Research on Similarity Recommendation Algorithm in Heterogeneous Information Network**

**ABSTRACT**

Recommendation is an effective way to address information overload and is widely used in e-commerce and Internet services. As an important data mining task, the recommendation system includes many object types and rich interactions between objects, which can be naturally modeled using the Heterogeneous Information Network (HIN). The superior information integration capability and rich semantic information of HIN further empower more exact recommendation. However, the existing HIN recommendation algorithms mainly focus on the interaction information between users and items, that is, purchase history. And this can only reflect one aspect of users’ taste and the characteristics of items in practical applications. Therefore, in order to dig deeper into the semantic similarity of various aspects, the Neural network based Aspect-level Collaborative Filtering model (NeuACF) designed a recommendation algorithm based on the meta-path, which makes the recommendation model fuse multiple aspects of semantic information and obtain more comprehensive and accurate results.

This paper first introduces the background of the topic and related technologies, and reviews the classic algorithms in the recommendation field from a macro perspective. Secondly, it analyzes the specific model structure of NeuACF in depth, and proposes the corresponding improvement schemes for its two problems. In particular, the possibilities of interaction between users and items utilize a dot product metric that may result in the destruction of the transfer characteristics of the similarity. To this end, this paper introduces metric learning, using the metric defined by the distance to capture the relationship between data, to achieve a more comprehensive information transmission. At the same time, NeuACF uses a point-wise loss function to highlight the absolute value of the rating information. However, for Top-N recommendations, the relative interaction possibilities are more important. Therefore, this paper considers using a pair-wise loss function based on negative sampling to produce a more reasonable list of recommendation. Finally, this paper has carried out extensive comparative experiments with the classical algorithms such as NeuACF on the recommended datasets Amazon and MovieLens, and verifies the superiority of the proposed model.

**KEY WORDS** recommendation heterogeneous information network metric learning

**目 录**

[第一章 引言 1](#_Toc8320190)

[1.1 课题背景 1](#_Toc8320191)

[1.1.1 研究意义 1](#_Toc8320192)

[1.1.2 研究现状 1](#_Toc8320193)

[1.2 课题任务 2](#_Toc8320194)

[1.2.1 课题内容 2](#_Toc8320195)

[1.2.2 本人承担任务 2](#_Toc8320196)

[1.3 论文结构 3](#_Toc8320197)

[第二章 相关技术介绍 5](#_Toc8320198)

[2.1 异质信息网络 5](#_Toc8320199)

[2.1.1 基本定义 5](#_Toc8320200)

[2.1.2 元路径 6](#_Toc8320201)

[2.2 注意力机制 7](#_Toc8320202)

[2.3 度量学习 10](#_Toc8320203)

[**2.3.1 形式化定义** 10](#_Toc8320204)

[**2.3.2 kNN的度量学习** 11](#_Toc8320205)

[2.4 本章小结 11](#_Toc8320206)

[第三章 基于异质信息网络的推荐模型详解 13](#_Toc8320207)

[3.1 基础模型 13](#_Toc8320208)

[**3.1.1 模型框架** 13](#_Toc8320209)

[**3.1.2 方面级相似性矩阵计算** 14](#_Toc8320210)

[**3.1.3 方面级隐因子学习** 14](#_Toc8320211)

[**3.1.4 基于注意力机制的方面级隐因子融合** 15](#_Toc8320212)

[**3.1.5 模型优化** 15](#_Toc8320213)

[3.2 问题分析 16](#_Toc8320214)

[3.3 改进方案 17](#_Toc8320215)

[3.4 本章小结 18](#_Toc8320216)

[第四章 实验 19](#_Toc8320217)

[4.1 实验设置 19](#_Toc8320218)

[4.1.1 数据集 19](#_Toc8320219)

[4.1.2 评价指标 19](#_Toc8320220)

[4.1.3 Baselines 19](#_Toc8320221)

[4.1.4 实现细节 19](#_Toc8320222)

[4.2 实验结果 19](#_Toc8320223)

[4.3 本章小结 19](#_Toc8320224)

[第五章 结束语 21](#_Toc8320225)

[N.1 论文工作总结 21](#_Toc8320226)

[N.2 问题和展望 21](#_Toc8320227)

[参考文献 22](#_Toc8320228)

[致 谢 24](#_Toc8320229)

[附 录 25](#_Toc8320230)

[外 文 资 料 26](#_Toc8320231)

[外 文 译 文 27](#_Toc8320232)

# 引言

## 1.1 课题背景

### 1.1.1 研究意义

推荐是解决信息过载的有效方法，被广泛应用于电子商务和互联网服务中。作为重要的数据挖掘任务，推荐系统包括许多对象类型（如电影推荐中的用户、电影、演员和导演等）及对象间的丰富交互关系（如观看信息、社交信息和属性信息等）。因此，融合更多信息进行混合推荐是解决数据稀疏性的有效技术。

近年来，大量涌现的社会媒体网站包含许多不同类型对象间的复杂交互，将这些对象建模为同质网络是很困难的。作为半结构化的表示方法，异质信息网络（Heterogeneous Information Network），即多种类型节点和链路的建模方式，可以有效建模和处理复杂多样的数据。其次，与同质网络相比，异质网络可以融合更多类型的对象及其复杂交互关系，也可以融合多个社交网络平台的信息。最后，异质信息网络包含丰富的语义，将导致更细微的知识发现。因此，异质信息网络作为有效的信息融合方法可以用于整合推荐系统中的对象和关系，可能产生更加准确的推荐结果。

### 1.1.2 研究现状

传统的推荐系统通常仅利用用户-物品评分反馈信息进行推荐。协同过滤[1]是最普及的技术之一，其基本假设是：有着相似购买历史的用户会倾向于购买相似的物品。为捕捉相似用户的购买偏好，隐因子模型（如矩阵分解）[2,3]应运而生，并在推荐系统中表现出令人瞩目的性能。具体地，它将用户-物品评分矩阵分解为两个低秩的用户因子和物品因子，然后利用隐因子进行预测。由于隐因子模型极易受数据稀疏性影响，随着社交媒体的普及，越来越多的研究人员通过整合辅助信息扩展隐因子模型，如异质网络推荐系统[4]。

近年来，异质信息网络，即多种类型节点和链路的建模方式，在数据挖掘领域大放异彩。作为重要的数据挖掘任务，推荐系统包括许多对象类型（如电影推荐中的用户、电影、演员和导演等）及对象间的丰富交互关系（如观看信息、社交信息和属性信息等），可以自然而然地利用异质信息网络建模。而异质信息网络优越的信息综合能力及丰富的语义信息使得更准确的推荐成为可能。

元路径（meta-path），即连接两类对象的关系组合，被广泛用于异质信息网络中的语义信息建模和关系抽取。Shi等人[5]实现了一个基于语义的推荐系统HeteRecom，利用元路径的语义信息来评估电影间的相似性。此外，考虑属性值，如链接上的评分，他们进一步将推荐系统建模为加权异质信息网络，并提出基于语义路径的个性化推荐方法SemRec[6]。为充分利用关系异质性，Yu等人[7]引入基于元路径的隐特征来表示沿不同类型路径的用户和物品间的连接，然后利用贝叶斯排名优化技术在全局和个性化水平上定义推荐模型。同样基于元路径，NeuACF[8]利用深度神经网络模拟和融合不同方面级别的隐因子，从不同方面表示用户偏好和物品特征，并自动确定其重要性。

许多方法利用异质信息网络来融合各种信息。结合不同上下文，Jamali等人[9]提出基于上下文的矩阵分解模型，考虑每个实体的一般隐因子和依赖于上下文的隐因子。使用用户隐式反馈数据，Yu等人[7]进一步解决全局和个性化实体推荐问题。根据相关的兴趣组，Ren等人[10]提出基于簇的引用推荐框架来预测每个查询在书目网络中的引用。同样，Wu等人[11]结合兴趣组信息，利用图表摘要和基于内容的聚类方法来进行媒体推荐。此外，利用多种类型的关系，Luo等人[12]提出一种社交协同过滤算法。

总的来说，大多数异质信息网络的相似性推荐算法集中在简单的网络结构（如二分或星型模式）和粗糙的语义建模（如元路径）上。但很多实际数据，特别是电子商务数据，所构建的异质信息网络包含很多节点和边类型，没有简单的网络模式。这种无模式异质信息网络带来了很多新的研究问题，如多种类型对象关系的管理及元路径的自动产生等，与此同时也亟需更强大的语义建模和分析方法挖掘丰富细致的语义信息。

## 1.2 课题任务

### 1.2.1 课题内容

本课题的主要内容是改进基于异质信息网络的推荐算法，使得推荐模型更好地融合多种语义信息，得到更加全面准确的结果。

为此，首先阅读异质信息网络综述，了解异质信息网络中的关键理论及相似性推荐算法的研究现状。建立了宏观概念后，有选择地调研相关论文，具体学习部分相似性推荐算法。在此基础上，对比分析现有的相似性推荐算法，选取合适的算法并改进。最终，完成相应改进模型的编码工作，并进行实验验证及结果评估。

### 1.2.2 本人承担任务

首先，了解课题背景及意义，调研相似性推荐算法，完成开题工作。

其次，基于对现有算法的分析，选择NeuACF作为基本模型，并剖析其存在的问题。具体地，NeuACF中，用户和物品的交互可能性是利用embedding间的点积来度量的。由于点积本身所固有的缺陷——不满足三角不等式，可能导致相似性的传递特征被破坏，从而限制模型效果的提升。与此同时，NeuACF使用point-wise损失函数，关注于评分信息的绝对数值。但对于Top-N推荐而言，相对的交互可能性更为重要。因而通过修改损失函数显式捕捉排序信息，可能产生更为精准的推荐列表。

针对分析所得结论，进一步完成相应改进方案的详细设计。具体地，引入metric learning，利用距离定义的metric捕捉数据间关系，弥补点积的不足，从而实现更全方位的信息传递。与此同时，使用基于负采样的pair-wise损失函数替换point-wise损失函数，产生排序更合理的推荐列表。

确定改进方案后，熟悉数据特点及相应算法框架。编码实现模型，完成中期报告及答辩。

最终，基于经典数据集进行实验验证，并对推荐结果进行分析评估，完成并提交毕业设计论文、相关文档及源程序，进行答辩。

## 1.3 论文结构

本文除引言外，后续内容可分为四个章节：

第二章主要介绍模型相关的背景技术，包括异质信息网络的关键概念、注意力机制的核心内容及度量学习的基础知识。

第三章主要介绍模型的详细方案。首先，阐述NeuACF的具体框架。其次，分析NeuACF存在的问题与不足，及其可能对模型性能产生的不良影响。最后，介绍如何结合先进方法改进NeuACF，及其背后相应的理论原因。

第四章主要介绍在经典数据集Amazon和MovieLens上所进行的对比实验，并对所得推荐结果进行解释分析，总结模型的优点及问题。

第五章，作为最后一个章节，对本文进行回顾，综合评价模型的创新点及进一步的优化方向，并对异质信息网络推荐算法的未来发展进行展望。

# 第二章 相关技术介绍

## 2.1 异质信息网络

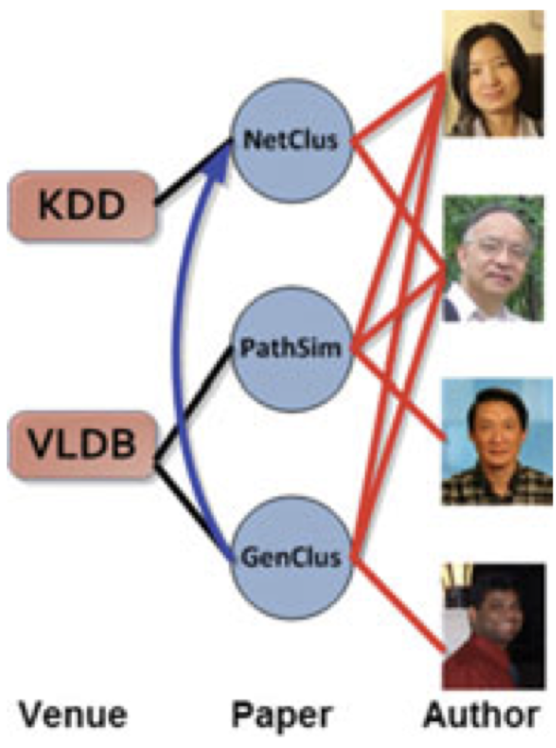
### 2.1.1 基本定义

众所周知，大多数实际系统通常由许许多多交互的、多类别的部件组成，如社交活动、计算机系统和生物网络等。在这些系统中，交互的部件构成信息网络。而无处不在的信息网络，形成了现代信息基础设施的重要组成部分，从而吸引了研究人员的广泛关注。一个信息网络代表现实世界中对象及其之间交互的抽象，正式定义如下：

**定义2.1**（信息网络[13,14])：信息网络定义为一个有向图,并具有对象类型的映射函数和链路类型的映射函数。每个对象属于对象类型集合中的某特定对象类型，每个链路属于关系类型集合中的某特定关系类型。如果两链路属于同一关系类型，那么这两个链路的始末对象类型分别相同。

为明确区分信息网络中的对象和关系类型，提出了异质/同质信息网络的概念，定义如下：

**定义2.2**（异质/同质信息网络）：若信息网络的对象类型数或关系类型数，那么称为异质信息网络；否则，称为同质信息网络。

（a）网络实例 （b）网络架构

图2-1 文献数据的异质信息网络[19]

图2-1是文献数据的异质信息网络示例[13]。一个文献信息网络，如从DBLP1中抽取的涉及计算机科学研究人员的文献网络，是典型的异质网络，它包括三种类型的实体：论文、会议和作者。对于每篇论文，都有到相应作者集合和会议的链路，这些链路属于链路类型的集合。

为更好地理解复杂异质信息网络中的对象和链路类型，有必要提供元层次（即模式层次）的网络描述。因此，进一步以网络模式来描述网络元结构：

**定义2.3**（网络模式[13, 14]）：网络模式记为，是信息网络的元结构，并有对象类型映射和链路类型映射。网络模式是定义在对象类型集合上的有向图，以上的关系为边。

异质信息网络的网络模式强调了对象集合与对象间关系类型的限定，这些限定使异质信息网络半结构化，并指导网络的语义探索。遵循某网络模式的信息网络称为网络模式的实例。对于从对象类型到对象类型的链路类型，即，和分别是链路类型的源对象类型和目标对象类型，可分别记为和。自然地，有逆关系。通常，不等于，除非是对称的。

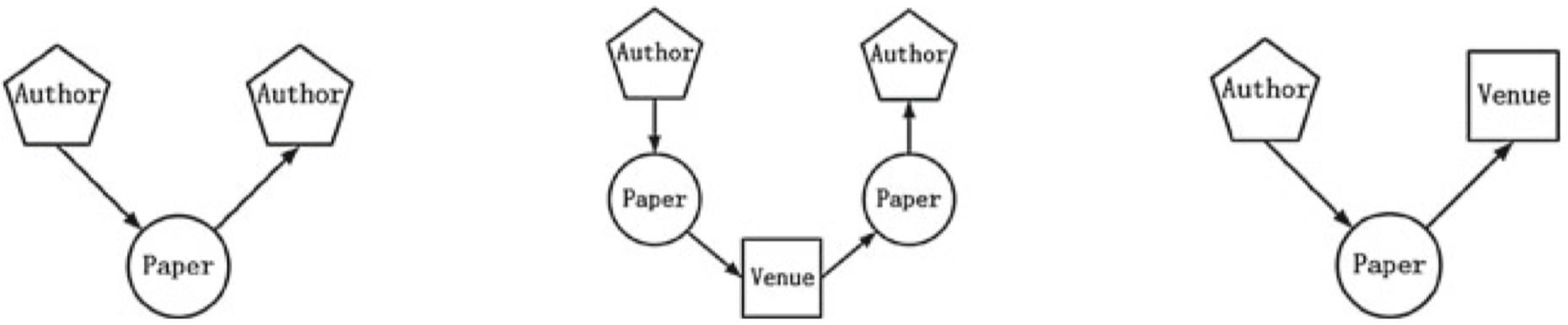
如上所述，图2-1（a）展示了文献数据中真实对象及其之间的联系，图2-1（b）展示了文献数据相应网络模式，且图2-1（a）是图2-1（b）的网络实例。在此示例中，包含三种类型对象：论文（），作者（）和会议（）。链路连接不同类型的对象，由两种对象类型间的关系定义。 例如，作者和论文间的链路表示写或被写关系，而会议和论文间的链路表示发表或发表于关系。

### 2.1.2 元路径

与同质网络不同，异质网络中两对象可通过不同的路径连接，这些路径具有不同的实际意义，从而以元路径分类：

**定义2.4**（元路径[15]）：元路径是在模式上定义的路径，记为。同时，定义对象间的复合关系，其中表示关系上的合成运算。

简单起见，若同一组对象类型间没有多种关系类型，可以使用对象类型表示元路径：。例如，图2-1（a）中作者在会议上发表论文，可以利用长度为2的元路径，或简写来描述。类似地，对象和间的具体路径是相关路径的路径实例。如果在中，对于每个，都有，且每个链路都属于关系，则记。当由元路径定义的关系是对称时（即等于），元路径是对称路径，如和。



(a) APA (b) APVPA (c) APV

图2-2 文献数据的元路径示例[19]

|  |  |  |
| --- | --- | --- |
| 路径实例 | 元路径 | 实际意义 |
| 孙-NetClus-韩 | 作者-论文-作者 | 作者合作同一篇论文 |
| 孙-PathSim-VLDB-PathSim-韩  孙-PathSim-VLDB-GenClus-阿加沃尔 | 作者-论文-会议-论文-作者 | 作者在同一会议发表论文 |
| 孙-NetClus-KDD  孙-PathSim-VLDB | 作者-论文-会议 | 作者在一个会议上发表论文 |

表2-1 关于文献数据的元路径示例及其实际意义

以图2-2所示的元路径为例，作者可以通过相应元路径连接，如“作者-论文-作者”路径，“作者-论文-会议-论文-作者”路径等。此外，表2-1说明了元路径的路径实例和相应语义。显然，这些路径的语义不同。路径指作者合作了一篇论文（即合作作者关系），而路径指作者在同一会议发表论文。

元路径包含丰富语义是异质网络的一个重要特征。基于不同的元路径，对象具有不同的连接关系和相应的路径语义，这可能会对数据挖掘任务的结果产生影响。例如，基于不同元路径评估作者间的相似性得分是不同的[15]。根据路径，共同发表论文的作者更相似，而根据路径，在同一会议发表论文的作者将更相似。作为一个独特的特征和有效的语义捕捉工具，元路径已被广泛使用在异质网络的信息挖掘中，如相似性度量[16, 15]、聚类[17]和分类[18]等。

## 2.2 注意力机制

在机器翻译领域，原始的方法大都利用数理统计进行分析和处理。近些年来，随着AlphaGo的出现，除了在游戏的人工智能领域，深度学习在计算机视觉和自然语言处理等领域也展现出了强大的效用。随着深度学习的进一步发展，Seq2Seq的训练和翻译模式开始进入人们的视野。除此之外，改进端到端的网络结构，弥补其已有的缺陷也是非常必要的。在此情形下，基于循环神经网络（Recurrent Neural Network,RNN）的注意力机制[20]被提出。因其显著的效果提升和端到端的高效训练方式，席卷人工智能各界，并快速应用至机器学习众多任务中。本文同样利用了注意力机制融合模型模块，因此简要介绍注意力机制的核心理论。

机器翻译，即输入某种语言中的一句话，输出目标语言相对应的语句。例如，将德语中的一段话翻译成合适的英语。之前的机器翻译模型通常是encoder-decoder结构，即encoder读取输入句子，将其转换为定长向量，然后decoder再将这个向量翻译成目标语言的对应文字。通常encoder及decoder均采用RNN结构，如长短期记忆网络（Long Short-Term Memory，LSTM）或门控循环单元（Gated Recurrent Unit，GRU）等。如图2-3所示，利用encoder将输入语句编码为最后一层中的隐向量，并将其作为decoder初始的隐向量，最终解码成目标语言中的文字。

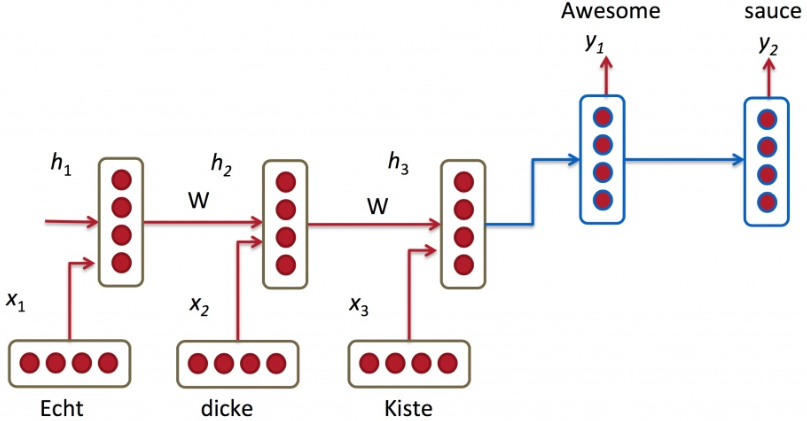


图2-3 encoder-decoder结构示例

但这个结构存在问题一些问题。其中，较为突出的问题是RNN机制中存在长程梯度消失，即对于较长的句子，很难将输入序列转化为定长向量而保存所有有效信息。因此，随所需翻译句子的长度的增加，这种结构的效果会显著下降。

而注意力机制，正是解决由长序列到定长向量转化而造成的信息损失问题。通俗理解，注意力机制跟人类自身翻译文章的思路类似，即将注意力放于翻译部分对应的上下文。相应地，注意力模型中，当翻译当前词语时，会寻找源语句中相对应的几个词语，并结合之前的已翻译部分得出后续翻译。如图2-4所示，当翻译“knowledge”时，只需将注意力放在源语句中“知识”的部分，当翻译“power”时，只需将注意力集中在“力量”部分。这样，当decoder预测目标翻译时，就可以考虑到encoder的所有信息，而不仅局限于原来模型中定长的隐藏向量，因此不会丧失长程信息。

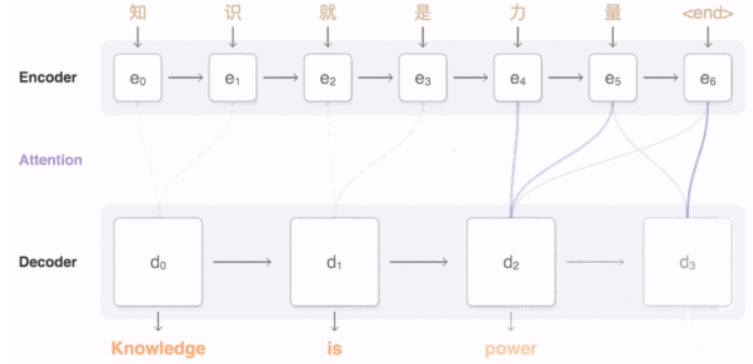


图2-4 基于注意力机制的机器翻译示例

以上是直观理解，接下来详细解释注意力机制在数学上对应哪些运算。

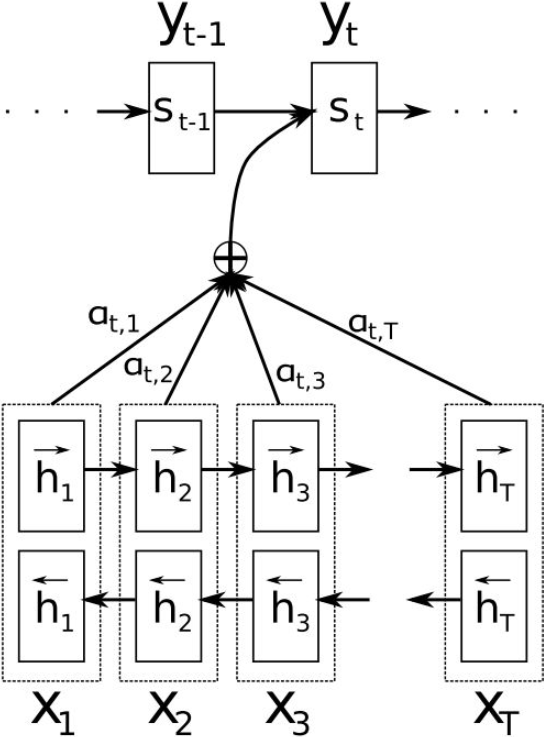


图2-5 注意力机制详解

首先，利用RNN结构得到encoder的隐状态。假设当前decoder的隐状态是，可以计算每一个输入位置与当前输出位置的关联性，相应向量形式为：

其中，是一种相关性运算符。例如，常见的有点乘形式、加权点乘、加和等。

然后，通过softmax将归一，得到注意力分布：

展开形式为：

利用可以加权求和得到相应的上下文向量：

由此，可以计算decoder的下一个隐状态：

以及该位置的输出：

上述过程的关键操作是计算encoder与decoder隐状态之间的关联性权重，得到注意力的分布。从而，对于当前输出位置比较重要的输入位置，相应地在预测时会占较大的比重。

通过注意力机制的引入，打破了只能利用encoder最终的单一向量的限制，使得模型可以集中在所有对于下一个目标单词重要的输入信息上，极大地改善模型效果。除此之外，通过观察注意力权重矩阵的变化，可以更好地知道哪部分翻译对应哪部分原文字，给模型提供更好的解释性，如图2-5所示：

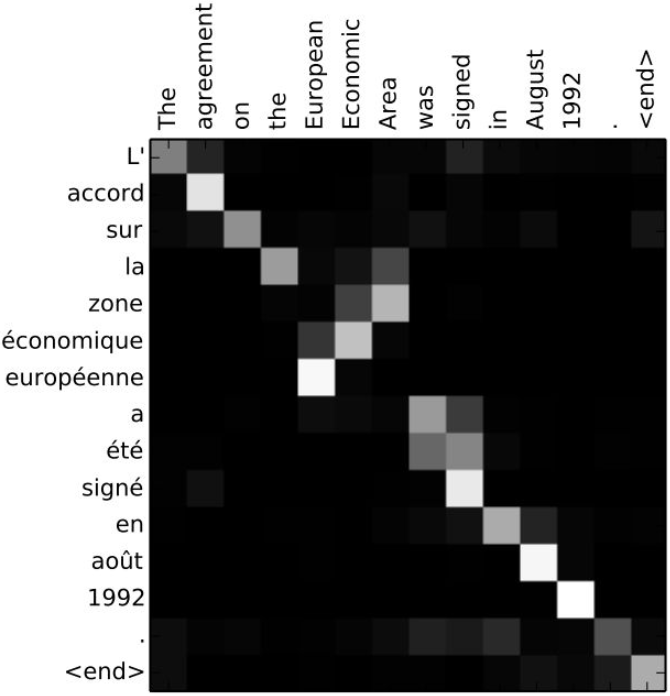


图2-5 注意力分布示例

## 2.3 度量学习

引入度量学习，是模型改进的出发点。因此，本节概述度量学习的基础知识，从而便于后续改进模型的阐述与理解。

距离，是许多机器学习算法的核心概念，包括K-nearest neighbors（kNN）、K-means和SVM等。度量学习，产生以距离为关键的度量指标，从而捕获数据间的关系。在许多机器学习任务中得到应用，包括图像分类和文档检索等[21,22,23,24]。

具体地，给定对象集合，已知某些对象是“相似的”或“不相似的”，度量学习的目标是学习一种符合相似和不相似关系的距离度量，即构建指标使得相似对间距离较小，不相似对间距离较大。

从数学的角度而言，度量指标需要满足几个条件，而三角不等式是其中最重要的一个[25]。具有三角不等式性质的度量指标可以表明，对于任意三个对象，任何两个成对距离的总和应大于或等于剩余的成对距离。这意味着，给定信息“x与y和z都相似”，度量指标不仅会拉近上述的两对关系，还会将剩余的关系对拉近。因此，可以将此视为相似性的传播过程，即度量指标将已知的相似性信息传播到关系未知的对象间。

**2.3.1 形式化定义**

设是输入空间上的数据集合。度量学习中的标签信息以成对的形式体现，包括已知相似对集合：

和已知不相似对集合：

基础的度量学习方法，构建以马氏距离为核心的度量指标：

其中，是半正定矩阵[25]。也就是说，将输入投影到新的空间，且在新空间中的欧氏距离服从给定的相似性约束。有许多种方法来构建这样的度量指标，最基础的做法是解决以下凸优化问题：

现有的先进方法倾向于使用非线性变换，如核函数或神经网络，来提高度量指标精度[26，27，28，29]。

**2.3.2 kNN的度量学习**

上述全局优化方法的本质是将所有相似对拉近，而将不相似对分开。一方面，这种优化并非始终可行。另一方面，Weinberger等人研究表明，若度量学习用于kNN，那么只需一个度量指标就可以让每个对象的kNN与该对象共享同类标签[27]，这个目标更易实现。本文中的改进模型也采用这种度量学习的概念，因为推荐的目标同样是找到每个用户的kNN物品。

具体地，给定输入，最接近的数据点称为的目标邻居。为了便于理解，可以认为的目标邻居建立了其余不同标记的输入不应入侵的边界。假设将侵入边界的不同标记的输入称为入侵者。那么，此时度量学习的目标就是构建一种最小化入侵者数量的指标[27]。

最经典的此类模型是large margin nearest neighbor（LMNN）[27]，它使用两个损失函数完成上述目标。即LMNN定义拉损失函数，将输入的目标邻居拉近：

其中，表示是的目标邻居。此外，LMNN进一步定义推损失函数，推动入侵者离开目标邻居，并在kNN决策边界周围保持较大距离：

如果和属于同一标签，则，否则，且是标准铰链损失。在此基础上，LMNN的总体损失函数是和的加权求和。

## 2.4 本章小结

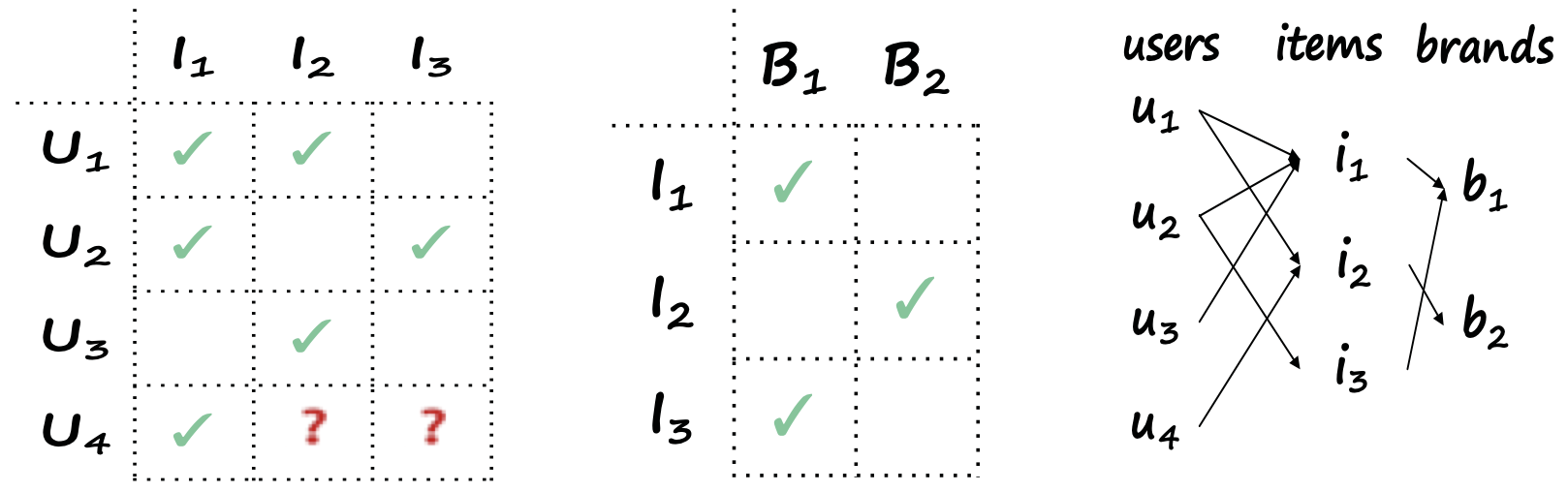
本章主要介绍了模型所依赖的相关背景技术。首先，解释了异质信息网络的基本定义及其关键语义挖掘工具——元路径。其次，阐述了注意力机制的核心理论。最后，说明了度量学习的相关概念，为后续模型的提出做好铺垫。

# 第三章 基于异质信息网络的推荐模型详解

## 3.1 基础模型

经广泛调研后，本文选择NeuACF[8]作为基础模型，并加以改进。因此，本节具体介绍NeuACF的基本思想及模型结构。

虽然现有的隐因子模型在推荐领域表现出了强劲性能，但它们通常只挖掘购买历史的信息，而这仅从一个方面反映用户偏好和物品特征。但是，用户和物品的隐因子源于实际应用的不同方面。这些方面，从不同角度反映用户偏好和物品特征。因此，隐因子模型应利用来自不同方面的信息。以图3-1为例，进一步说明多方面信息的重要性。如果仅利用图3-1（a）中的交互矩阵（即购买历史），会推断出用户将购买物品和。然而，若考虑到图3-1（b）所示的品牌信息时，可以发现物品可能更受用户喜爱，因为物品和属于同一品牌。



（a）用户-物品交互矩阵 （b）物品-品牌关系矩阵 （c）异质信息网络

图3-1 用户-物品方面级交互关系示例[8]

尽管全面利用各方面信息是极有价值的，但面临两个严峻挑战。其一，如何提取不同方面级的信息。其二，如何从不同方面学习和融合隐因子。即使能够提取出不同方面的特征，学习相应的隐因子并实现有效融合仍不是一件易事。虽然可以利用矩阵分解学习隐因子，但它只能学习“浅层”的隐因子。

为解决上述挑战，基于深度神经网络的NeuACF应运而生，可以有效提取和融合不同方面级的隐因子。具体地，推荐系统中不同类型的对象和其之间的交互被建模为异质信息网络，并利用元路径[15]从不同方面提取用户和物品特征。如图3-1（c）所示，可以通过用户-物品-用户路径，从购买历史角度提取用户的隐特征。此外，进一步利用深度神经网络捕捉复杂的非线性关系，为用户和物品学习不同方面级的隐因子，并基于注意力机制实现有效融合，进行Top-N推荐。

**3.1.1 模型框架**

NeuACF模型包含三个主要步骤。首先，基于推荐系统中丰富的用户-物品交互信息构造异质信息网络，并在不同元路径下计算不同方面级的相似性矩阵，以反映用户和物品的不同方面级特征。接下来，利用深度神经网络，以所得相似性矩阵为输入，分别学习方面级隐因子。最后，结合注意力机制融合方面级隐因子，获得用户和物品的最终表示。接下来，详细说明以上三个步骤。

**3.1.2 方面级相似性矩阵计算**

给定元路径，选择相似性矩阵提取方面级特征。原因有二，其一，相似性度量可以减弱噪声；其二，相似性数值在之间，便于隐因子学习。因此，使用PathSim[15]计算方面级相似性矩阵，如基于元路径User-Item-Brand-Item-User（）和Item-Brand-Item（）可以提取品牌方面特征。

**3.1.3 方面级隐因子学习**

已计算出不同方面级用户-用户和物品-物品相似性矩阵，接下来需要利用它们学习相应隐因子，架构如图3-2所示：

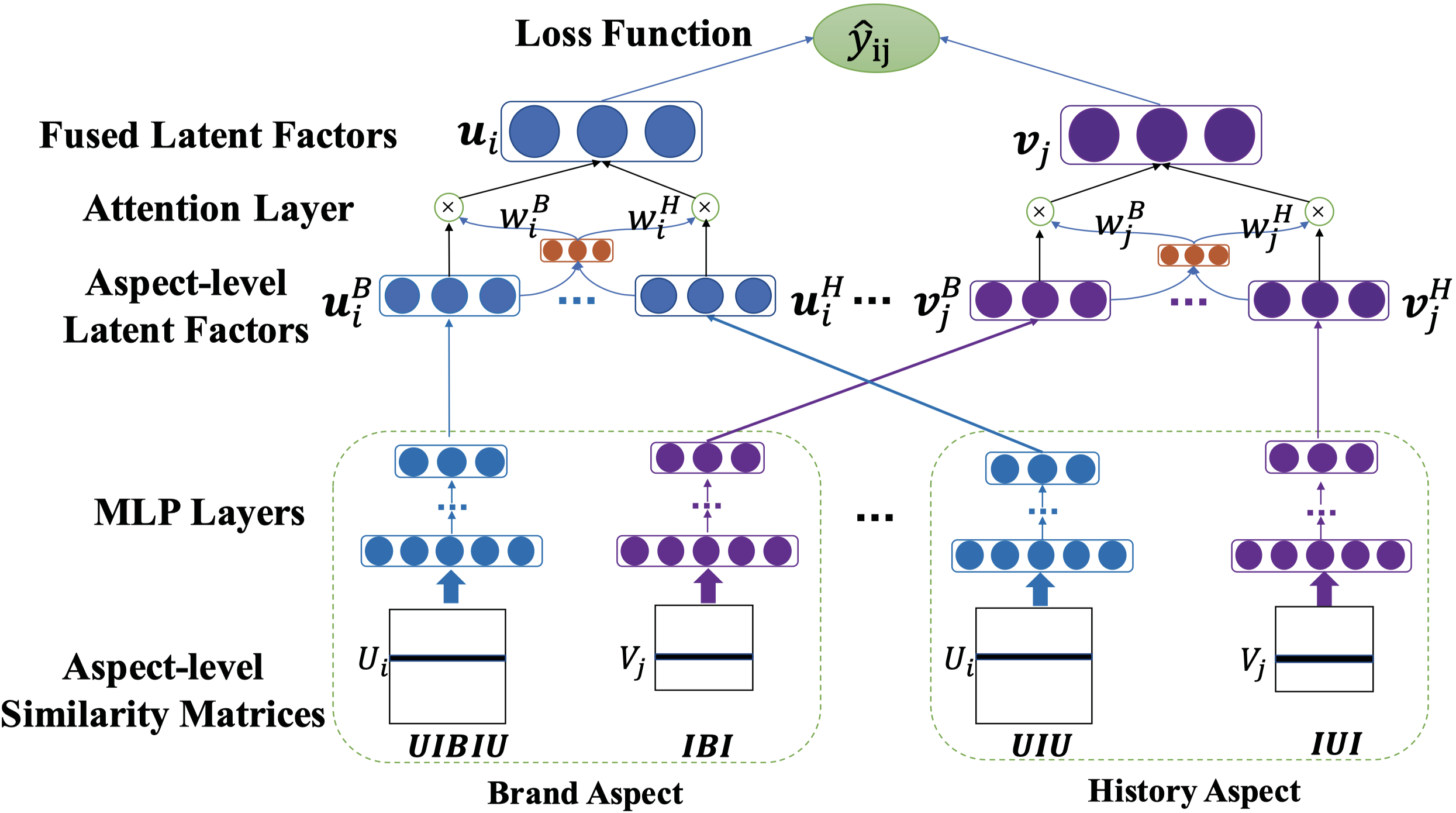


图3-2 NeuACF中的深度神经网络架构

具体地，以方面级相似性矩阵作为多层感知器（Multi-Layer Perceptron，MLP）的输入。相应地，方面级隐因子为其输出。以元路径下的用户相似性矩阵为例，用户以维的向量表示，该向量反映用户与所有其他用户的相似性。其中，是用户总数。MLP将用户的初始相似性向量投影到低维方面级隐因子。在MLP的每层中，输入向量被投影为新空间中的一个向量。形式化如下，给定初始输入向量和隐藏层，其中代表第层，可以通过以下多层映射来学习最终的方面级隐因子：

其中，和分别是第层的权重矩阵和偏置，是隐藏层的激活函数。从图2-4中的架构可以看出，对于用户和物品的每个方面级相似性矩阵，均存在相应的MLP学习该方面级隐因子。

**3.1.4 基于注意力机制的方面级隐因子融合**

已获得用户和物品的方面级隐因子后，接下来需要将它们融合形成最终的表示。比较简单的方法是直接拼接所有方面级隐因子，或者平均所有方面级隐因子。但是，这两种方法都没有区分不同方面的重要性。这显然是不合理的，因为并非所有方面对于推荐都是有意义的。因此，考虑利用先进的注意力机制融合方面级隐因子。

注意力机制在多种机器学习任务中都表现出了优越的特性，如机器翻译[30]。基于注意力机制，可以为所有方面级隐因子分配合适的权重：较高（较低）权重反映相应方面对于推荐是信息量充足（信息量匮乏）的。具体地，利用用户的品牌方面隐因子，通过双层注意力网络计算相应权重：

其中，是权重矩阵，是偏置。

利用Softmax函数归一化，从而获得各方面级隐因子的最终权重，即方面对用户最终表示的贡献：

其中，是方面集合。

获得用户的所有方面级隐因子的权重后，可以计算最终表示：

**3.1.5 模型优化**

将Top-N推荐建模为分类问题，该问题预测用户与物品间的交互概率。为确保输出值是概率，需要将输出约束在内。使用Logistic函数作为输出层的激活函数，根据下式计算用户和物品间的交互概率：

其中，和分别是用户和物品的最终表示。

在所有训练集上，似然函数是：

其中，和分别是正负实例集合，是参数集合。

利用似然函数的负对数形式，得到point-wise损失函数：

其中，是实例的ground truth，是预测结果。上式即为模型的目标函数，可以通过随机梯度下降或其变种技术来优化。

## 3.2 问题分析

NeuACF中，用户和物品的交互可能性是利用embedding间的点积度量的。由于点积本身所固有的缺陷——不满足三角不等式，可能导致相似性的传递特征被破坏，从而限制模型效果的提升。与此同时，NeuACF使用point-wise损失函数，关注于评分信息的绝对数值。但对于Top-N推荐而言，相对交互可能性更为重要。接下来，本节具体分析以上两点问题可能对于模型效果产生的不良影响。



图3-3 矩阵分解稳定解示意图

首先，如图3-3所示，、和是三个规模相同的用户组。其中，喜欢物品,喜欢物品,喜欢物品和。图3-3左半部分展示了在矩阵分解框架下的一组稳定解，即当用户喜欢物品时，用户和物品间的点积等于2，否则点积等于0。可以观察到，虽然用户同时喜欢物品和，但是物品和间的点积等于0。这一点违反了三角不等式，即相似对和间的相似关系无法传播到。

这种相似性传播被打破的情况，可能造成两种后果。第一，尽管矩阵分解框架在两个轴上捕获了最突出的隐因子，但是并没有反映用户的反馈信息中包含的更细粒度偏好。第二，隐因子不能准确捕获用户-用户或物品-物品间的相似性，而这对于缓解协同过滤的数据稀疏性问题至关重要。

其次，point-wise损失函数作为预测结果与ground truth的交叉熵，更为关注分数的绝对数值。而这种损失函数形式对于推荐任务而言，存在两点不足。第一，推荐产生的结果往往是一个物品列表，即Top-N推荐，因此追求的是排序的准确性，而非精确打分数值，只需相对打分即可。第二，point-wise损失函数没有建模预测物品列表中的排序信息，可能会无意中过分强调那些不重要的物品，即排序结果后部，用户并不喜爱的物品。因此，Top-N推荐效果可能受到限制。

## 3.3 改进方案

针对已分析的两点问题，本节阐述具体的改进方案及相应理论基础。

首先，相似性的传递与协同过滤技术密切相关。因为，在协同过滤中，也是通过已知的某些用户-物品对间关系，来推断其他未知关系的用户-物品对。正如上节所分析的那样，因三角不等式的性质未被满足，使用用户和物品间embedding的点积来捕捉已有的评分信息，并非是一种度量学习方式。从而，导致相似性的传递过程可能被打破。

为此，引入距离的概念。距离，是许多机器学习算法的核心概念，如K-nearest neighbor（kNN）、K-means等。度量学习利用距离定义的度量指标捕捉数据间的关系，从而使得相似对间距离较近，不相似对间距离较远。而度量学习所需满足的重要性质之一——三角不等式，自然地蕴含相似性的传递过程，即将已知的相似性信息通过距离传递给未知的点对，这与协同过滤的本质思想不谋而合。因此，本文考虑以用户-物品度量指标替换NeuACF中的点积估计形式，在反映用户偏好的同时，捕捉用户-用户及物品-物品间的相似性，实现更全方位的信息传递。具体地，对于已知相似关系的用户-物品对集合，度量指标将会使得中相应对的embedding距离较近，而不在中的点对距离较远。由于满足三角不等式，在此过程中会使得喜爱相同物品的用户聚成簇，且被相同用户喜爱的物品聚成簇。最终，对于用户的kNN物品将是：该用户曾喜爱的物品或具有相同口味的相似用户曾喜爱的物品。换句话说，通过学习满足已知相似信息的度量指标，不仅可以将这些信息传播到其他未知的用户-物品对，而且还可以传播给无法直接观察的用户-用户和物品-物品对。

形式化而言，将用户的embedding记为，物品的embedding记为。以它们之间的欧几里得距离作为度量指标，修改NeuACF打分函数：

从而满足用户对不同物品的相对喜爱程度，即喜爱的物品将比不喜爱的物品更接近该用户。

其次，既然已经在模型中引入了度量学习，那么相应地需要使用适合于度量学习的损失函数，即pair-wise损失函数。其输入是一个三元组，是目标对象，是正样本，即同类对象，是负样本，即异类对象。对应于本文的推荐任务，也就是，是用户，是用户喜爱的物品，是用户不喜爱或不清楚喜好关系的物品。有如下不等式：

其中，表示某种度量指标下的距离数值，threshold表示特定阈值。该不等式本质上定义了正负样本之间的距离关系，即正样本间的距离与阈值之和会小于负样本间的距离。因此，基于此种三元组输入和不等式理论，相应的pair-wise损失函数为：

其中，是正样本集合。因此，最终的优化目标是拉近正样本的距离，推远负样本(i, k)的距离。显然，只有括号内公式的值大于0时，才可以计算误差，并依据反向传播调节网络参数。

## 3.4 本章小结

本章详细介绍了模型各模块的组成。首先，阐述了基础模型NeuACF的具体框架。其次，分析了NeuACF存在的两点问题，即点积和point-wise度量方式可能对模型性能造成的不良后果。最后，引入度量学习和pair-wise损失函数的理论，阐明改进NeuACF的方案，形成最终的模型结构。

# 第四章 实验

## 4.1 实验设置

### 4.1.1 数据集

本文利用三个公开的数据集评价模型性能，即亚马逊数据集和不同规模的MovieLens数据集。数据集的网络模式如图4-1所示，具体统计信息如表4-1所示。

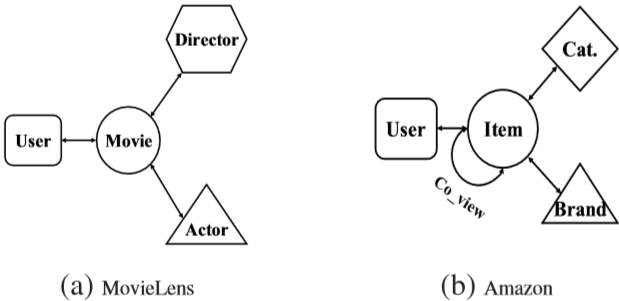


图4-1 实验数据集的网络模式

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Datasets** | **#users** | **#items** | **#ratings** | **#density** |
| **ML100K** | 943 | 1682 | 100,000 | 6.304% |
| **Amazon** | 3532 | 3105 | 57,104 | 0.521% |

表4-1 实验数据集的统计信息

MovieLens-100K（ML100K）2：MovieLens数据集广泛用于电影推荐中。本文使用ML100K版本。为构建异质信息网络，需要对原有数据集进行处理，即对于每部电影，从IMDb中爬取相应导演和演员。

Amazon3：该数据集包含亚马逊电商网站中的用户评分数据。在实验中，本文选择电子类商品进行模型评估。

公开数据集MovieLens和Amazon所包含的用户-物品交互信息为1至5的相应评分，因而直接建模会得到带权异质信息网络。但就本文所提模型而言，不考虑权重对于欧几里得距离度量指标的影响，因此将其统一处理为1。也就是说，若存在评分信息，相应用户-物品有连边；若不存在评分信息，则相应用户-物品无连边。

### 4.1.2 评价指标

本文采用leave-one-out法对模型性能进行评估。具体地，选择一个用户已评分的物品作为正样本，随机选择99个未被该用户评分的物品作为负样本。对于该用户而言，模型将对此100个物品进行排序，产生相应的推荐列表。为与baselines进行公平比较，本文对所有baselines测试集中每个（用户，物品）正样本，使用相同的负样本集合。

在此基础之上，通过命中率（Hit Ratio ，HR）和归一化折损累计增益（Normalized Discounted Cumulative Gain，NDCG）作为具体评价指标。

对于HR而言：

其中，是对测试集中所有用户来说，产生的相应推荐列表中出现正样本的用户。

而NDCG通常用于评价搜索算法，后续用于推荐算法的评价。首先介绍其基本思想，即高关联度的结果比一般关联度的结果更影响最终的指标数值，且高关联度的结果出现在列表更靠前的位置时，指标数值会越高。具体地，先从累计增益（Cumulative Gain，CG）开始理解。CG只考虑了关联度，而没有考虑到所处列表位置的影响。它实际上是搜索结果相关性分数的总和：

其中，代表位置上搜索结果的相关性分数。

进一步，出现了CG的改进版本DCG（Discounted Cumulative Gain），即在每个CG的数值上除以一个折损值。其目的在于，让排名越靠前的结果越影响最后的评价指标，换句话说，排序越靠后，价值越低。对于位置上的搜索结果，其价值是，那么产生的效益就为：

但是，由于检索词的不同，返回的搜索结果数量是不一致的。而DCG是一个累加值，无法针对两种不同的搜索结果进行比较。因此，需要归一化处理，最终出现了NDCG：

其中，IDCG表示理想情况下的最大DCG值。

实际应用到本文所采用的leave-one-out法，NDCG的具体形式为：

其中，是正样本在推荐列表中所处的排序位置。

与此同时，由于效率原因，本文并不产生全部的物品推荐列表，而是利用Top-K排序结果。其中，。

### 4.1.3 Baselines

本文共利用8个基线进行对比实验，包括基础模型NeuACF、两个基本方法（ItemPop和ItemKNN）、两个矩阵分解方法（Matrix Factorization，MF[2]和element-wise Alternating Least Squares，eALS[31]）、一个基于pair-wise损失函数的排序方法（Bayesian Personalized Ranking，BPR[32]）和两个基于神经网络的方法（Deep Matrix Factorization Models, DMF[33]和Neural networkbased Collaborative Filtering，NCF[34]），概要介绍如下：

NeuACF，将推荐系统建模为异质信息网络，利用元路径捕捉用户和物品不同方面级隐因子，并实现高效融合。

ItemPop，物品按照受欢迎程度，即与所有用户的交互总次数，排序推荐。

ItemKNN，基于物品的标准协同过滤方法。

MF, 矩阵分解框架中极具代表性的隐因子模型。

eALS，用于推荐的矩阵分解算法，基于element-wise交替最小二乘技术，有效优化具有可变加权缺失数据的矩阵分解模型。

BPR，利用贝叶斯后验估计，实现个性化排序，并进一步使用pair-wise损失函数优化矩阵分解模型。

DMF，利用深度神经网络的矩阵分解模型，将用户和物品映射到共同的低维向量空间。

NCF，基于神经网络的协同过滤算法，结合了矩阵分解的线性和深度神经网络的非线性，用于对用户-物品隐结构进行建模。

### 4.1.4 实现细节

本文基于TensorFlow深度学习框架实现所提模型，并对所有数据集，调节相应超参数，使得模型达到最佳效果。对于神经网络，MLP层数在集合中调节，每个隐藏层具有600个隐藏单元。潜在因子维数在集合中调节。batch大小在集合中调节，学习率在集合中调节。在训练模型时，对于每个正样本，负采样的负样本数量在集合中调节,且度量中的margin在集合中调节。模型用Xavier初始化器随机初始化模型参数，并使用Adam作为优化器优化模型。表4-2说明了实验中所用的元路径及相应方面。与此同时，对于所有baselines，根据相应文献，设定其最佳参数。

|  |  |  |  |
| --- | --- | --- | --- |
| **数据集** | **方面** | **元路径** | |
| **用户** | **电影/物品** |
| **MovieLens** | 历史 | UMU | MUM |
| 导演 | UMDMU | MDM |
| 演员 | UMAMU | MAM |
| **Amazon** | 历史 | UIU | IUI |
| 品牌 | UIBIU | IBI |
| 种类 | UICIU | ICI |
| 共同浏览 | UIVIU | IVI |

表4-2 实验所用元路径及其对应方面

## 4.2 实验结果

### 4.2.1 推荐性能分析

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Datasets | Metrics | ItemPop | ItemKNN | MF | eALS | BPR | DMF | NCF | NeuACF | Ours |
| ML100K | HR@5 | 0.2831 | 0.4072 | 0.4634 | 0.4698 | 0.4984 | 0.3483 | 0.4942 | 0.5097 | 0.3648 |
| NDCG@5 | 0.1892 | 0.2667 | 0.3021 | 0.3201 | 0.3315 | 0.2287 | 0.3357 | 0.3505 | 0.2351 |
| HR@10 | 0.3998 | 0.5891 | 0.6437 | 0.6638 | 0.6914 | 0.4994 | 0.6766 | 0.6846 | 0.5483 |
| NDCG@10 | 0.2264 | 0.3283 | 0.3605 | 0.3819 | 0.3933 | 0.2769 | 0.3945 | 0.4068 | 0.2915 |
| HR@15 | 0.5366 | 0.7094 | 0.7338 | 0.7529 | 0.7741 | 0.5873 | 0.7635 | 0.7813 | 0.6713 |
| NDCG@15 | 0.2624 | 0.3576 | 0.3843 | 0.4056 | 0.4149 | 0.3002 | 0.4175 | 0.3844 | 0.3249 |
| HR@20 | 0.6225 | 0.7656 | 0.8144 | 0.8155 | 0.8388 | 0.6519 | 0.8324 | 0.8464 | 0.7572 |
| NDCG@20 | 0.2826 | 0.3708 | 0.4034 | 0.4204 | 0.4302 | 0.3151 | 0.4338 | 0.4002 | 0.3442 |
| Amazon | HR@5 | 0.2412 | 0.1897 | 0.3027 | 0.3063 | 0.3296 | 0.2693 | 0.3117 | 0.3216 | 0.3395 |
| NDCG@5 | 0.1642 | 0.1279 | 0.2068 | 0.2049 | 0.2254 | 0.1848 | 0.2141 | 0.2232 | 0.2472 |
| HR@10 | 0.3576 | 0.3126 | 0.4278 | 0.4287 | 0.4657 | 0.3715 | 0.4309 | 0.4686 | 0.4613 |
| NDCG@10 | 0.2016 | 0.1672 | 0.2471 | 0.2441 | 0.2693 | 0.2179 | 0.2524 | 0.2683 | 0.2863 |
| HR@15 | 0.4408 | 0.3901 | 0.5054 | 0.5065 | 0.5467 | 0.4328 | 0.5258 | 0.5591 | 0.5298 |
| NDCG@15 | 0.2236 | 0.1877 | 0.2676 | 0.2647 | 0.2908 | 0.2332 | 0.2774 | 0.2924 | 0.3045 |
| HR@20 | 0.4997 | 0.4431 | 0.5680 | 0.5702 | 0.6141 | 0.4850 | 0.5897 | 0.6257 | 0.5895 |
| NDCG@20 | 0.2375 | 0.2002 | 0.2824 | 0.2797 | 0.3067 | 0.2458 | 0.2925 | 0.3080 | 0.3186 |

表4-3 不同方法的HR@K和NDCG@K指标对比

表4-3展示了不同方法的实验结果。可以看出，本文所提模型并非在所有数据集和指标上都能达到最佳性能。具体地，相比较Amazon数据集，ML100K数据集更加稠密，评分信息较多。因其数据特点，baselines中基于评分信息的方法都有较好的表现

NeuACF几乎在所有数据集和标准上都达到了最佳性能。 作为具有神经网络的最新模型，NeuMF在大多数情况下也表现良好，而NeuACF在几乎所有情况下都始终优于NeuMF，只有一个例外。 我们认为原因在于NeuACF学到的潜在因素的多个方面提供了用户和项目的更多整体特征。 虽然FMGals利用与NeuACF相同的特征，NeuACF的更好性能意味着NeuACF中的深度神经网络可能比FMG中的“阴影”模型更好地学习用户和项目的潜在因素。

### 4.2.2 margin大小对性能的影响

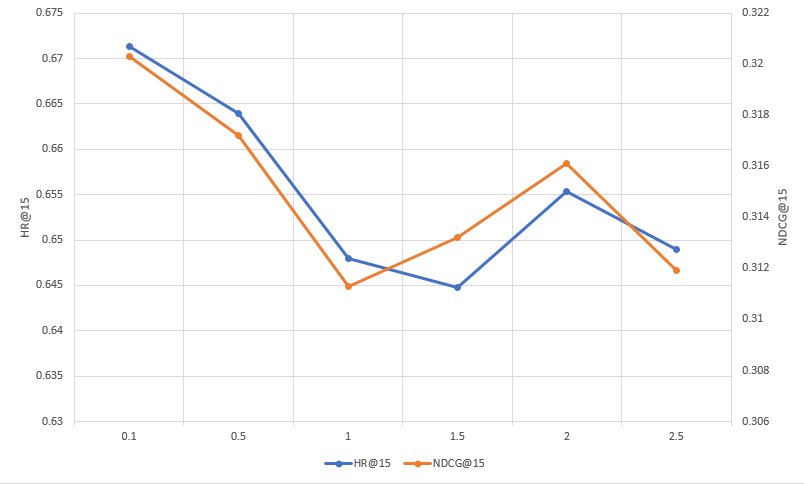


图4-2 不同margin对应的推荐性能

在基于欧几里何距离的度量指标中，margin的大小，可能会对推荐性能产生重要影响。因此，本文具体研究了所提模型的margin大小变动时，推荐性能的相应变化趋势。由于超参数量较多，本文采取控制变量法，即固定MLP层数为2，batch大小为1024，隐因子数量为32，学习率为0.00005，负采样数量为10，使得margin大小从0.1至2.5变动。因时间限制，本文选择在ML100K数据集上进行实验，其他数据集上的结论应近似，结果如图4-2所示。可以看出，随margin的增加，推荐性能虽稍有波动，但总体呈下降趋势。究其原因，可能是在数据集规模较大的情况下，大的margin可能导致收敛速度降低，甚至难以收敛。

### 4.2.3 负采样数量对性能的影响

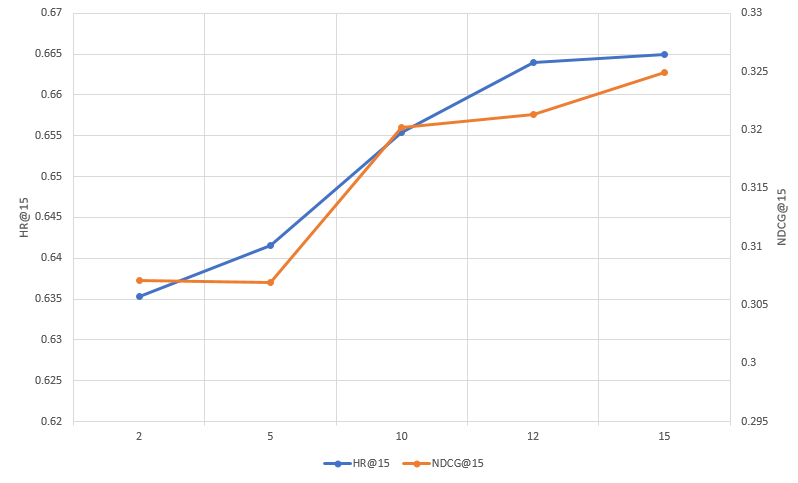


图4-3 不同负采样数量对应的推荐性能

在基于负采样技术的pair-wise损失函数中，负采样的数量，可能对推荐性能产生重要影响。因此，本文具体研究了所提模型的负采样数量变动时，推荐性能的相应变化趋势。由于超参数量较多，本文采取控制变量法，即固定MLP层数为2，batch大小为1024，隐因子数量为32，学习率为0.00005，margin大小为2.0，使得负采样数量从2至15变动。因时间限制，本文选择在ML100K数据集上进行实验，其他数据集上的结论应近似，结果如图4-3所示。可以看出，随负采样数量增加，推荐性能逐渐增加，并慢慢趋于稳定。因此，为了平衡性能与效率，负采样数量可以在15左右选择。

### 4.2.4 隐因子数量对性能的影响

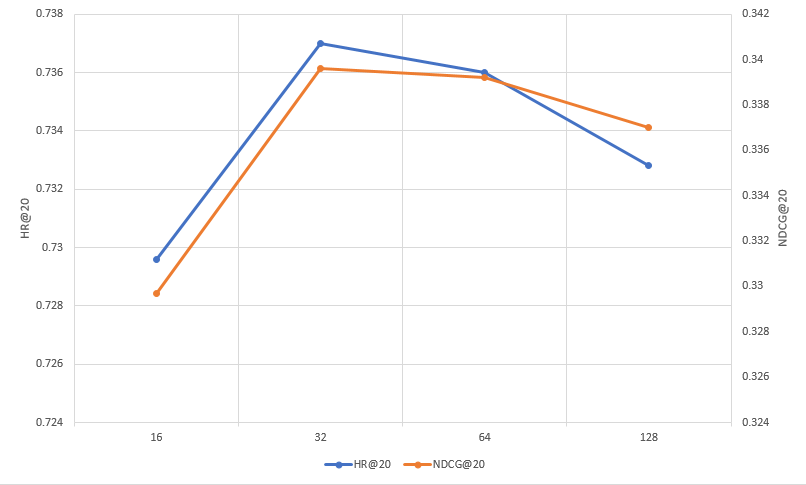


图4-4 不同隐因子数量对应的推荐性能

在隐因子模型中，隐因子的数量，即embedding维度，可能对推荐性能产生重要影响。因此，本文具体研究了所提模型在MLP最后一层所学到的隐因子的数量变动时，推荐性能的相应变化趋势。由于超参数量较多，本文采取控制变量法，即固定MLP层数为2，batch大小为1024，学习率为0.00005，负样本数量为10，margin大小为2.0，使得隐因子维数自16至256变动。因时间限制，本文选择在ML100K数据集上进行实验，其他数据集上的结论应近似，结果如图4-4所示。可以看出，随隐因子数量增加，推荐性能首先增加，峰值性能在32-64左右达到。此时，如果维度进一步增大，推荐性能会下降，可能产生了过拟合。

## 4.3 本章小结

# 第五章 结束语

## 5.1 论文工作总结

本文通过选题背景介绍，从宏观角度回顾了推荐领域的经典算法。进一步地，深入分析了基础模型NeuACF的具体结构，并论述了其存在的两点问题，给出相应改进方案。具体地，引入metric learning和pair-wise损失函数，产生排序更合理的推荐列表。最终通过编码实现改进模型，在推荐数据集Amazon和MovieLens上，与经典算法进行了大量对比实验，验证了所提模型的优越性。

综上，本文取得的主要成果有三点：

* 1. 在基础模型NeuACF中引入度量学习的概念，以欧几里得距离指标捕捉数据间关系，弥补点积所造成的相似性传递特征丢失，实现更全方位的信息传递。
  2. 使用pair-wise损失函数替换point-wise损失函数，强调对于Top-N推荐而言的相对交互可能性，产生排序更合理的物品推荐列表。
  3. 基于推荐数据集Amazon和MovieLens，与经典算法进行对比实验，验证了模型的优越性。

由于毕设内容是本人在石川老师指导下，独立完成的。因此，主要工作有：

1. 了解课题背景及意义，调研相似性推荐算法，完成开题。
2. 基于对现有算法的分析，选择NeuACF作为基本模型，并剖析其存在的问题。
3. 针对分析所得结论，进一步完成相应改进方案的详细设计。
4. 确定改进方案后，熟悉数据特点及相应算法框架。编码实现模型，进行中期检查。
5. 基于经典数据集进行实验验证，并对推荐结果进行分析评估，完成并提交毕业设计论文、相关文档及源程序。

## 5.2 问题和展望

### 5.2.1 问题

### 5.2.2 展望

随着大数据时代的来临，网络信息快速膨胀，高效且个性化推荐的重要性愈发凸显。因此，受到了学术界和产业界的联合关注，但不得不承认推荐技术目前仍面临诸多难点，如冷启动及多模态数据的融合运用等。

与此同时，异质信息网络作为一个年轻且快速发展的研究领域，可以自然地用于推荐系统的建模。因此，展望基于异质信息网络的推荐技术，可能有如下几个研究方向：

首先，是更复杂的异质信息网络构建。已有的相关研究，大多假定异质信息网络是明确定义的，即对象和对象间的关系是清晰的。然而，在实际推荐系统中，构造异质信息网络会遇到很多挑战。对于关系数据库等的结构化数据，构造相应异质信息网络较容易。然而即使是这种网络，对象和对象间的关系也可能具有噪声，如对象重名或关系不完整等。对于文本、图像等非结构化数据，如何准确抽取相应对象和关系，进而建立完善和准确的异质信息网络，也同样是难点。在实践中，会利用多种技术，如自然语言处理、图像处理等。

其次，是更强大的分析方法。在异质信息网络中，对象通过不同的方式组织在一起。星型模式是广泛使用的异质信息网络类型。在推荐中，网络化数据更加复杂且没有规律性。举个例子，网络中的链路会包含属性值，如评分信息，从而构成带权异质信息网络。这种丰富模式的异质网络，亟需强有力的语义分析工具。

最后，是更大数据的处理。为展现异质网络建模的优势，需要在更广泛的领域中设计实用的推荐算法。多样性是大数据的重要特征，而异质网络是处理多样性的有效方法。实际中，异质网络是巨大的，甚至是动态的，所以通常不能在内存中直接处理。由于用户往往只对一小部分节点、链路或子网络感兴趣，可以从现有网络中动态地提取子网络进行分析。

前路漫漫，但机遇与挑战并存。随着国内外学者持续的关注与投入，推荐技术必能够越来越成熟，为人类带来更好的生活体验。

# 参考文献

[1] Hu Y, Koren Y, Volinsky C. Collaborative Filtering for Implicit Feedback Datasets[C]//ICDM. 2008, 8: 263-272.

[2] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009 (8): 30-37.

[3] Koren Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 426-434.

[4] 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.

[5] Shi C, Zhou C, Kong X, et al. Heterecom: a semantic-based recommendation system in heterogeneous networks[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1552-1555.

[6] Shi C, Zhang Z, Ji Y, et al. SemRec: a personalized semantic recommendation method based on weighted heterogeneous information networks[J]. World Wide Web, 2019, 22(1): 153-184.

[7] 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. ACM, 2014: 283-292.

[8] Han X, Shi C, Wang S, et al. Aspect-Level Deep Collaborative Filtering via Heterogeneous Information Networks[C]//IJCAI. 2018: 3393-3399.

[9] Jamali M, Lakshmanan L. HeteroMF: recommendation in heterogeneous information networks using context dependent factor models[C]//Proceedings of the 22nd international conference on World Wide Web. ACM, 2013: 643-654.

[10] Ren X, Liu J, Yu X, et al. Cluscite: Effective citation recommendation by information network-based clustering[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 821-830.

[11] Wu J, Chen L, Yu Q, et al. Trust-aware media recommendation in heterogeneous social networks[J]. World Wide Web, 2015, 18(1): 139-157.

[12] 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.

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

[14] Sun Y, Yu Y, Han J. Ranking-based clustering of heterogeneous information networks with star network schema[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 797-806.

[15] Sun Y, Han J, Yan X, et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.

[16] Shi C, Kong X, Yu P S, et al. Relevance search in heterogeneous networks[C]//Proceedings of the 15th international conference on extending database technology. ACM, 2012: 180-191.

[17] Sun Y, Norick B, Han J, et al. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2013, 7(3): 11.

[18] Kong X, Yu P S, Ding Y, et al. Meta path-based collective classification in heterogeneous information networks[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012: 1567-1571.

[19] Shi C, Philip S Y. Heterogeneous information network analysis and applications[M]. Springer International Publishing, 2017.

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[21] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1701-1708.

[22] Wan J, Wang D, Hoi S C H, et al. Deep learning for content-based image retrieval: A comprehensive study[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 157-166.

[23] Koestinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2288-2295.

[24] Xu Z E, Chen M, Weinberger K Q, et al. From sBoW to dCoT marginalized encoders for text representation[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012: 1879-1884.

[25] Xing E P, Jordan M I, Russell S J, et al. Distance metric learning with application to clustering with side-information[C]//Advances in neural information processing systems. 2003: 521-528.

[26] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009, 10(Feb): 207-244.

[27] Wang J, Do H T, Woznica A, et al. Metric learning with multiple kernels[C]//Advances in neural information processing systems. 2011: 1170-1178.

[28] Cogswell M, Ahmed F, Girshick R, et al. Reducing overfitting in deep networks by decorrelating representations[J]. arXiv preprint arXiv:1511.06068, 2015.

[29] Kedem D, Tyree S, Sha F, et al. Non-linear metric learning[C]//Advances in Neural Information Processing Systems. 2012: 2573-2581.

[30] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[31] He X, Zhang H, Kan M Y, et al. Fast matrix factorization for online recommendation with implicit feedback[C]//Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2016: 549-558.

[32] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback[C]//Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence. AUAI Press, 2009: 452-461.

[33] Xue H J, Dai X, Zhang J, et al. Deep Matrix Factorization Models for Recommender Systems[C]//IJCAI. 2017: 3203-3209.

[34] He X, Liao L, Zhang H, et al. Neural collaborative filtering[C]//Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 173-182.

1 http://dblp.uni-trier.de/

2 https://grouplens.org/datasets/movielens/

3 http://jmcauley.ucsd.edu/data/amazon/

# 致 谢

行文至此，意味着本科的学习生活即将结束。四年的时间，说长不长，说短也不短，是旧时期的结束，也是新阶段的开始。在这匆匆的几年中，我有许多想要感谢的人，谢谢他们一路的鼓励与支持，谢谢他们如一的包容与理解，谢谢有人教我如何成长，谢谢有人教我怎样爱与被爱。

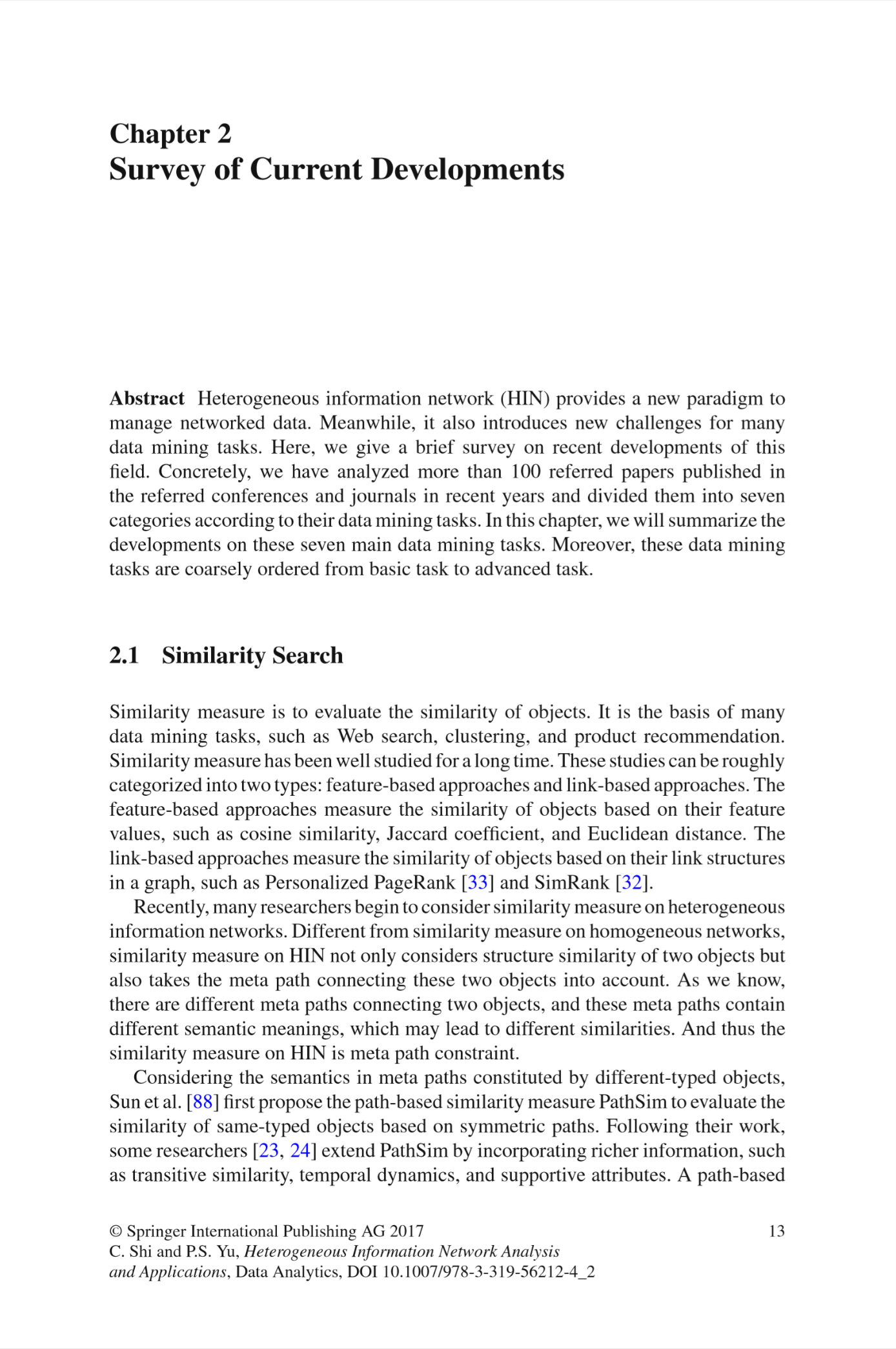
首先，我要感谢毕业设计的指导老师——石川老师，北京邮电大学计算机学院教授，也是我未来博士研究生的导师。作为我的研究生导师，石川教授在学术研究中认真严谨，对于我的毕业设计给予了很多建议与帮助，并及时指出其中存在的问题。与此同时，他也耐心引导我适应学术研究的过程，让我从最初的迷茫困惑，到现在渐渐明晰方向。

其次，我要感谢实验室的两位青年教师及师兄师姐。顺利保研后，经半年相处，我认为用大家庭来形容实验室更为准确。除了和师兄师姐的融洽交流，两位年轻老师亦师亦友，在这种和谐的环境中，无论是科研还是项目，都有事半功倍的效果。在此期间，我也逐渐明白了真正做科研应是怎样的心态与方式。

再次，我要感谢我的舍友、同学、我爱的人和爱我的人，他们为我的生活增添了许多明媚的色彩，使我从踌躇与痛苦中脱身而出，实现蜕变。

最后，感谢努力而不言败的自己。

# 外 文 资 料



# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 外 文 译 文

**第 2 章 发展现状概述**

**摘要**

异质信息网络提供了管理网络化数据的新模式。与此同时，它也为许多数据挖掘任务带来了新的挑战。在这里，我们简要介绍这个领域的最新发展。具体而言，我们分析了近年来相关研究工作的100多篇会议和期刊论文，并根据其针对的数据挖掘任务将这些工作分为七类。在本章中，我们将概述这七个主要数据挖掘任务的发展，这些任务按从基本到高级粗略排序。

**2.1 相似性搜索**

相似性度量用于评估对象的相似性，是许多数据挖掘任务的基础，如Web搜索，聚类和产品推荐等。就相似性度量已进行了长时间的研究，这些研究大致分为两类：基于特征和基于链接。基于特征的方法根据特征值来度量相似性，如余弦相似度，Jaccard系数和欧几里德距离。基于链接的方法根据图中的链接结构来度量相似性，如Personalized PageRank[33]和SimRank[32]。

最近，许多研究人员开始考虑异质信息网络中的相似性。与同质网络上的相似性度量不同，异质信息网络上的相似性度量不仅考虑两个对象间的结构相似性，还考虑两个对象相连的元路径。众所周知，连接两个对象的元路径不同，则包含的语义不同，从而可能导致不同的相似性。因此，异质信息网络的相似性度量是元路径约束的。

考虑到不同类型对象构成的元路径所包含语义，Sun等人首先提出PathSim方法来评估基于对称路径的相同类型对象相似性。之后，一些研究人员[23,24]通过整合更丰富的信息来扩展PathSim，如传递相似性、时间动态和支持属性等，进一步提出了基于路径的相似性连接方法[108]，根据用户指定的连接路径返回前k个相似对象对。Wang等人[101]定义基于元路径的关系相似性度量方法RelSim，用于度量丰富模式的异质信息网络中关系实例间的相似性。此外，Wang等人[99]将文档建模为异质信息网络，并提出新型相似性度量方法KnowSim来计算两个文档的相关性。在信息检索领域，Lao和Cohen[46,47]提出了一种路径约束的随机游走模型来度量由科学文献元数据构建的标记有向图中的实体接近度。

为了评估不同类型对象的相关性，Shi等人[72,74]提出HeteSim方法度量任意元路径下对象对的相关性。作为HeteSim的改进版本，LSH-HeteSim[48]用于在异质生物网络中挖掘药物靶标相互作用，其中药物和靶标与复杂的语义路径相关联。为克服HeteSim计算和内存性能的高需求缺点，Meng等人[62]提出AvgSim度量方法，分别通过沿给定元路径和反向元路径的两个随机游走过程来评估相似性。此外，一些方法[8,141]将基于元路径的相关性搜索与用户偏好相结合。

尽管基于元路径的相似性度量方法已经可以较有效地捕获源和目标对象间的单一关系，如在元路径APA下的共同作者，但在应用中仍然存在一些缺点，如书目数据中，希望根据以下事实来衡量两位作者的关系：他们的论文不仅发表在同一会议上，且具有相同的主题（即APVPA和APTPA路径）。为衡量对象间的复杂相关性，Huang等人[28]提出基于元结构的相关性度量，它是一种有向无环图，可以认为是元路径的组合。同样，Fang等人[20]将元图作为一种新颖的手段，用于表征给定接近类别的共同结构。此外，他们还提出了一系列基于元图的接近度，并采用监督技术自动学习正确的接近形式以适应给定类别。

更多的工作开始整合网络结构和其他信息，以度量异质信息网络中对象的相似性。结合影响和相似性信息，Wang等人[102]同时度量异质网络中的社交影响和对象相似性，从而产生更有意义的相似性分数。Wang等人[96]提出一个模型，通过分析异质网络的在线定位背景来学习相关性。Yu等人[116]基于元路径特征空间中的用户查询来预测语义含义，并训练排名模型用于回答相似性查询。最近，Zhang等人[133]提出一种新的相似性度量方法，用于根据属性相似性和中心间的连接来计算x-star网络中心间的相似性。

**2.2 聚类**

聚类分析是将数据对象划分为一组簇的过程，簇中的对象彼此相似，但与其他簇中的对象不同。常规聚类方法基于对象的特征，如k-means等[30]。近期，基于网络化数据（如，社团检测）的聚类方法激发了广泛的研究。此类方法将数据建模为同质网络，并使用给定度量（如，标准化切割[78]和模块化[63]）将网络划分为一系列子图，现已提出许多算法来解决这个NP-hard问题，如谱方法，贪婪方法[93]和采样技术[71]等。一些研究还同时考虑对象的链接结构和属性信息，以提高聚类准确性[110,140]。

最近，异质网络的聚类引起了很多关注。与同质网络相比，异质网络集成了多类型对象，从而为聚类带来了新的挑战。一方面，网络中共存的多类型对象会导致新的聚类范例，异质信息网络中的簇可能包含相同主题的不同类型对象[82]。如，在书目异质网络中，数据库区域的簇由一组作者、会议、术语和论文组成。通过这种方式，异质信息网络中的聚类可以保留更丰富的信息，但也面临更多困难。另一方面，异质信息网络中包含的丰富信息使得聚类中集成附加信息或其他学习任务更加方便。在本节中，我们将根据集成信息或任务的类型来回顾这些工作。

属性信息被广泛地使用在异质信息网络的聚类分析中。Aggarwal等人[1]使用局部的简洁属性在异质网络中创建平衡社团。考虑到异质信息网络中对象属性和类型链接的不完整性，Sun等人[85]提出一种基于模型的聚类算法，用于整合不完整的属性信息和网络结构信息。Qi等人[67]提出基于异质随机场的聚类算法，用于对具有异常链接的社交媒体网络的结构和内容进行建模。Cruz等人[14]整合结构维度和组成维度，构建归因图来解决社团检测问题。近期，基于密度的聚类模型TCSC[7]用于同时利用网络中连接和顶点属性的聚类。

文本信息在许多异质网络研究中起着重要作用。Deng等人[17]提出一个具有偏差传播的主题模型，以统一的方式将异质信息网络与主题建模相结合。此外，他们[16]还提出了一个联合概率主题模型，用于同时建模异质信息网络的多类型对象。LSA-PTM[103]通过不同对象间的链接在异质信息网络上传播，并利用LSA获得的主题来识别多类型对象的簇。Wang等人[104]将文档内容和各种链接结合在异质网络中，提出用于主题挖掘和多类型对象聚类的统一主题模型。近期，CHINC[98]使用通用知识作为间接监督来进一步改善聚类结果。

用户指南信息也可以集成到聚类分析中。Sun等人[87]提出了一种半监督聚类算法，用于根据用户指南生成具有路径选择的不同聚类结果。Luo等人[58]首先介绍相关路径的概念以及同类型对象间的相似性，并使用标记信息加权关系路径，然后提出SemiRPClus方法用于异质信息网络中的半监督学习。

聚类通常是一项独立的数据挖掘任务。但是，它可以与其他挖掘任务相结合，通过相互增强来提高性能。最近，基于排序的异质信息网络聚类方法已经出现，从而表明聚类和排序可以相互促进。RankClus方法[83]基于聚类和排名相互增强的想法，为双星网络中特定类型对象生成簇。NetClus方法[82]用于处理具有星型模式的网络。Wang等人[105]通过应用具有自循环的星型模式网络来组合异质和同质信息，并引入ComClus方法以提高聚类和排名的性能。此外，通用方法HeProjI通过将网络投影到子网序列，可以在任意模式的异质网络中进行基于排序的聚类[75]。Chen等人[12]提出一种概率生成模型，用于在任意模式的异质网络上同时实现聚类和排序。为利用文本信息和异质链接实体，Wang等人[97]提出一种聚类和排序算法，自动构建多层主题层次结构。更进一步，Qiu等人[68]提出OcdRank算法，将重叠社团检测和社团成员排名结合在定向的异质社交网络中。

异常值检测是使用与期望非常不同的行为来查找数据对象的过程。异常值检测和聚类分析是两个高度相关但不同的任务。为检测异常值，Gupta等人[21]提出基于联合非负矩阵分解的异常值识别方法，用于发现普及的社团分布模式。此外，他们提出在给定联合选择查询的情况下，检测异质网络中基于关联的clique异常值[22]。更进一步，Zhuang等人[142]提出一种异常检测算法，根据不同的查询和语义找到子网异常值。同样基于查询，Kuck等人[44]提出基于元路径的异常值度量方法，用于挖掘异质网络中的异常值。

此外，某些聚类方法还集成了一些其他信息。例如，基于社会影响的聚类框架SI-Cluster利用用户间的联系和他们的社交活动来分析异质信息网络[138]。除用于异质信息网络聚类的传统模型，如主题模型和谱聚类，Alqadah等人[4]提出了一种新的博弈论框架，用于在异质信息网络中定义和挖掘簇。

**2.3 分类**

分类是一种数据分析任务，通过构建模型或分类器来预测类标签。传统的机器学习主要针对满足独立同分布的相同结构对象的分类。但是，许多真实世界数据集中的对象间存在链接，使得其不满足独立同分布。因此，基于链接的对象分类受到广泛关注，其中数据图由通过一组链接彼此相连的对象构成。许多方法扩展了传统的分类方法，从而考虑对象间的相关性[45]。基于链接的对象分类通常认为图中的对象和链接是相同的，也就是说，构成同质网络。

与传统的分类研究不同，异质信息网络研究的分类问题具有一些新的特点。首先，异质信息网络中包含的对象是不同类型的，这意味着可以同时对多类型的对象进行分类。其次，标签可以通过不同类型对象间的各种链接传播。在异质信息网络中，对象的标签由其不同类型链接所相连的不同类型对象综合决定。

许多工作将传统分类方法扩展到异质信息网络。其中，一些工作扩展了转换分类任务，即预测给定未标记数据的标签。例如，GNetMine方法[35]用信息网络模拟具有任意网络模式和任意数量的对象/链接类型的链接结构。Wan等人[94]结合典型基于图形的转换分类方法和为同质网络而设计的转换回归模型，提出一个图形正则化的基于元路径的转换回归模型。Luo等人提出HetPathMine方法[56]，通过新的元路径选择模型在异质信息网络上聚集小标记数据。Jacob等人[29]提出通过计算空间中节点的潜在表示来标记不同类型的节点，其中，两个连接的节点具有相近的潜在表示。最近，Bangcharoensap等人[6]采用边缘中介中心性进行边缘权重归一化，并进一步提高中心性，使其适用于异质网络。另外，一些工作还扩展了归纳分类，即在整个数据空间中构建决策函数。例如，Rossi等人[70]使用二分异质网络来表示文本文档，并提出IMBHN算法来构建分配给文本术语相应权重的分类模型。

多标签分类在实际应用中很普遍，即每个例子可同时与一组标签相关联[41]，这种分类任务也扩展到了异质信息网络中。Angelova等人[5]提出一种基于多标签图的分类模型，通过将节点间的相互影响建模为随机游走过程来标记异质网络。Kong等人[41]使用从异质信息网络连接结构中挖掘出的多类型关系来促进多标签分类过程。Zhou等人[139]考虑结构亲和性和标签邻近性，提出以边缘为中心的多标签分类方法。

作为一种独特的特征，元路径广泛用于异质信息网络的分类。许多方法利用元路径生成特征，如GNetMine[35]和HetPathMine[56]。Kong等人[40]引入对象间基于元路径的依赖关系，用于研究集体分类问题。最近，Wang等人[100]利用文本构建的异质信息网络所包含的元路径，开发相应核方法用于文本分类。

与聚类问题类似，分类也可以与异质信息网络中的其他数据挖掘任务相集成。基于排名的分类是将分类和排名整合到一个同步且相互增强的过程中。Ji等人[36]提出一个基于排名的分类框架RankClass，从而执行更准确的分析。作为RankClass的延伸，Chen等人[13]提出统一分类框架F-RankClass，应用于单峰或多峰数据的二元或多类分类。一些方法还将分类与信息传播相结合。例如，Jendoubi等人[34]基于社交信息在网络中的传播和信念函数理论对社交信息进行分类。

**2.4 排名**

排名是网络分析中重要的数据挖掘任务，即基于一些排名函数来评估对象重要性或流行度。在同质网络中已提出了许多排名方法，如PageRank[65]和HITS[39]，这些方法仅考虑同质网络中的相同类型对象。

异质信息网络中的排名是一项重要而有意义的任务，但面临着一些挑战。首先，在异质信息网络中存在不同类型的对象和关系，若同等处理所有对象，会将不同类型的对象混合在一起。其次，异质信息网络中不同类型的对象和关系具有不同的语义含义，这可能导致不同的排名结果。以书目异质网络为例，在不同的元路径下，作者的排名会有不同的结果[50,77]，因为这些元路径将在作者间构建不同的链接结构。而且，不同类型对象的排名具有相互的影响。例如，享有声誉的作者通常会在顶级会议上发表论文。

二分图中的联合排名问题在过去几十年中激发了广泛的研究。例如，Zhou等人[137]通过耦合两个随机游走过程，联合排名作者及其出版物，而co-HITS方法[15]将二分图结合了内容信息和相关约束。Soulier等人[80]提出双重实体排名算法，通过组合基于内容和基于网络的特征，在主题查询的书目网络中联合排名文档和作者。在多关系网络上也有一些排名工作。例如，MultiRank方法[64]同时确定对象和关系对多关系数据的重要性，并且提出HAR 方法[49]来确定对象的中心和权限得分，用于给出查询搜索多关系数据中的相关性得分。这两种方法关注具有多关系的相同类型对象。最近，Huang等人[26]将正式文体和推断的社交网络与Twitter相结合，对推文进行排名。虽然这项工作在异质网络中使用了各类型的对象，但它仍然只针对一种类型的对象排名。

考虑到异质信息网络上元路径的特征，一些工作提出了基于路径的排名方法。例如，Liu等人[54]通过利用异质书目图上的多个元路径，开发具有伪相关反馈的出版物排名方法。利用张量分析，Li等人[50]提出HRank方法同时评估多类型对象和元路径的重要性。

排名问题也扩展到了社交媒体构建的异质信息网络中。对于社交媒体中的图像搜索，Tsai等人[91]提出SocialRank方法，使用社交提示进行图像搜索和社交网络排名。为识别Q＆A系统中的高质量对象（问题，答案和用户），Zhang等人[129]提出一个无监督的异质网络框架，从而在Q＆A站点中对多个对象进行联合排名。对于异质跨域排名问题，Wang等人[95]提出通用的正则化框架来发现两个域的潜在空间，并在潜在空间中同时最小化两个加权排名函数。考虑到文学网络的动态性，一个新的相互强化排名框架被提出，可以同时对新出版物和年轻研究人员的未来人气进行排名[106]。

**2.5 链路预测**

链路预测是链路挖掘中的基本问题，即尝试基于观察到的链路和节点属性来估计两节点间存在链接的可能性。链路预测通常被视为简单的二进制分类问题：对于任何两个可能链接的对象，预测链路存在（1）或不存在（0）。一种方法是完全基于网络的结构属性进行预测[51]，另一种方法是利用属性信息进行链路预测[66]。

近年来，异质信息网络中的链路预测一直是一个重要的研究课题，具有以下特征：首先，要预测的链接具有不同类型，因为异质信息网络中的对象与不同类型的链接相连。其次，在多种类型的链接间存在依赖性。因此，异质信息网络中的链路预测需要捕获不同类型链接间存在的不同且复杂的关系，并利用补充预测信息来共同预测多种类型的链接。

利用元路径，许多工作采用两步来解决异质信息网络中的链路预测问题。第一步，提取基于元路径的特征向量。第二步，训练回归或分类模型来计算链接的存在概率[10,11,84,86,115]。例如，Sun等人[84]提出PathPredict方法，通过结合基于元路径的特征提取和基于逻辑回归的模型来解决共同作者关系的预测。Zhang等人[130]基于元路径来预测组织结构图或管理层次结构。Cao等人[10]利用各种复杂的联系信息，设计相关性度量方法构建链接的特征向量，并提出迭代框架来共同预测多种类型的链接。此外，Sun等人[86]利用提取的拓扑特征，模拟关系建立时间的分布，从而预测何时形成某种关系。

概率模型也广泛应用于异质信息网络中的链路预测任务。Yang等人[112]提出一种概率方法MRIP，模拟在异质关系间的传播，从而预测多关系异质网络中的链路。此外，TFGM模型[113]定义潜在的主题层来桥接多个网络，并构建半监督学习模型来挖掘异质网络间的竞争关系。Dong等人[19]开发基于转移的排名因子图模型，将几种社交模式与网络结构信息相结合，用于链路预测和推荐。矩阵分解是处理链路预测问题的另一种常用工具。例如，Huang等人[27]开发联合流形分解方法（JMF），通过聚合异质社交网络，利用辅助评级矩阵进行可信预测。

上述方法主要关注单个异质网络上的链路预测。最近，Zhang等人[42,126,128]提出跨多个对齐的异质网络的链路预测问题，在[42]中提出一种两阶段链路预测方法。第一阶段，从多个网络中提取异质特征，第二阶段通过将锚链接公式化为稳定的匹配问题来推断锚链接。另外，Zhang等人[126]提出SCAN-PS方法，使用“锚点”解决新用户的社交链路预测问题。此外，他们还提出TRAIL[128]方法同时预测社交链路和位置链路。针对新用户的冷启动问题，Liu等人[52]通过利用来自另一个类似的社交网络的信息，提出用户-用户链路预测的对齐因子图模型。为识别来自多个异质社交网络的用户，并集成不同的网络，在考虑多网络间的本地和全局一致性的基础上，提出基于能量的模型COSNET[134]。

关于链路预测的大多数工作都是针对静态网络而设计的，然而，动态链路预测也非常重要且具有挑战性。考虑到Web数据的动态和异质性，Zhao等人[135]提出一个通用框架，用于从异质Web数据的演变中表征和预测社团成员。为解决时间和异质信息网络中的动态链接推断问题，Aggarwal等人[2,3]提出一个两级方案，可以有效地进行宏观和微观决策，从而组合拓扑和类型信息。Ma等人[60]提出利用两个新结构的进化因子模型，即邻居分布向量和邻居标签进化矩阵，目的是预测标签在给定节点的邻居上的分布。

**2.6 推荐**

推荐系统帮助消费者搜索可能感兴趣的产品，如书籍、电影和餐馆等，利用信息检索、统计和机器学习的各种技术来搜索产品和客户偏好间的相似性。传统的推荐系统通常仅利用用户产品评分反馈信息进行推荐。协同过滤是最普及的技术之一，包括两种方法：基于记忆的方法和基于模型的方法。最近，矩阵分解体现出其在推荐系统中的有效性和效率，它将用户产品评分矩阵分解为两个低级的用户-特征和产品-特征矩阵，然后利用分解矩阵进行预测[81]。随着社交媒体的普及，越来越多的研究人员通过利用用户间的社交关系[59,111]研制社交推荐系统。

最近，一些研究人员开始意识到异质信息对于推荐的重要性。异质信息网络全面的信息和丰富的语义使其有望产生更好的推荐结果。例如，电影推荐系统[76]构建的异质信息网络中，不仅包含不同类型的对象（如，用户和电影），而且还说明了对象间的各种关系，如观看记录、社交关系和属性信息等。构建异质网络进行推荐可以有效地融合各种信息，而且网络中的对象和关系具有不同的语义，从而利于揭示对象间的微妙关系。

元路径可以很好地探索语义并提取对象间的关系。Shi等人[73]实现了一个基于语义的推荐系统HeteRecom，利用元路径的语义信息来评估电影间的相似性。此外，考虑属性值，如链接上的评分，他们进一步将推荐系统建模为加权异质信息网络，并提出基于语义路径的个性化推荐方法SemRec[76]。为充分利用关系异质性，Yu等人[117,118]引入基于元路径的潜在特征来表示沿不同类型路径的用户和产品间的连接，然后利用贝叶斯排名优化技术在全局和个性化水平上定义推荐模型。同样基于元路径，Burke等人[9]提出在加权混合中包含多个关系的推荐方法。

许多方法利用异质信息网络来融合各种信息。结合不同的上下文，Jamali等人[31]提出基于上下文的矩阵分解模型，考虑每个实体的一般潜在因子和依赖于上下文的潜在因子。使用用户隐式反馈数据，Yu等人[117,118]进一步解决全局和个性化实体推荐问题。根据相关的兴趣组，Ren等人[69]提出基于簇的引用推荐框架来预测每个查询在书目网络中的引用。同样，Wu等人[107]结合兴趣组信息，利用图表摘要和基于内容的聚类方法来进行媒体推荐。基于多种异质网络特征，Yang等人[109]使用基于SVMRank的方法将多个特征建模至统一的框架中。利用多种类型的关系，Luo等人[57]提出一种社交协同过滤算法。此外，采用用户和产品的相似性作为正规化，一些工作[75,136]提出用于推荐的矩阵分解框架。

**2.7 信息融合**

信息融合，即从具有不同概念、上下文和印刷表示的异质源合并信息的过程。由于各种数据源的可用性，融合这些分散的分布式信息源已成为一个重要的研究问题。过去的几十年中，在许多传统数据挖掘领域已发表了数十篇关于该主题的论文，如数据仓库中的数据模式集成[61]、蛋白质-蛋白质相互作用网络（PPI）、生物信息学中匹配的基因调控网络[79]和Web语义中的本体映射[18]等。如今，随着异质信息网络数量的激增，跨多个异质信息网络的信息融合已成为一个新颖但意义重大的研究问题。通过融合不同异质信息网络，可以获得在不同异质信息网络中共享公共信息的实体更全面和更一致的知识，包括结构、属性和活动等。

为融合多个异质信息网络，一个重要的前提条件是通过共享公共信息的实体来对齐异质信息网络，实体可以是社交网络中的用户、书目网络中的作者和生物网络中的蛋白质分子等。完美的异质信息网络对齐是一个具有挑战性的问题，因为潜在的子图同构问题实际上是NP-complete[38]的。同时，基于异质信息网络中可用的结构和属性信息，目前为止已提出了大量异质信息网络的近似对齐算法。受[92]中的同质网络对齐方法启发，Koutra等人[43]利用快速网络对齐算法将两个二分图对齐。Zafarani等人[138]基于各种节点属性（如，用户名、打字模式和语言模式等）跨社交网络匹配用户。Kong等人[42]将异质社交网络对齐问题表述为锚链接预测问题，在[42]中提出一种两步监督方法MNA，用于异质信息网络上的潜在锚链路。然而，现实世界中的大部分社交网络实际上是部分对齐的，且许多用户不是锚用户。针对这种情况，Zhang等人[123,131]分别提出基于监督学习设置和PU学习设置的部分网络对齐方法。除这些成对的社交网络对齐问题外，还可以同时对齐多个（两个以上）社交网络。Zhang等人[124]发现社交网络对齐中实体的跨网络映射应满足传递性法则，且具有固有的一对一约束，在[124]中引入一个新的多元社交对齐框架，以最小化对齐成本并保持传递法则和一对一约束。除用户外，许多其他类型的实体也可由多个社交站点共享，如社交网络共享的地理空间位置和电子商务站点共享的产品等。为同时推断这些不同类型实体间的映射关系，Zhang等人提出[125]中的网络部分联合对齐问题。

通过融合多个异质信息网络，每个网络中的异质信息可转移到其他对齐网络，从而使得异质信息网络上的许多应用问题，如链路预测、朋友推荐[90,123,127]、社团检测[122]、信息扩散[119,120,132]和产品推荐[55]等，将从中受益匪浅。

通过映射，Zhang等人[123,127]提出跨越对齐网络传输的异质链路，以提高预测链接/推荐朋友的质量。Tang等人[90]提出基于转移的因子图模型，通过借用不同源网络的知识来预测目标网络中的社交关系类型。对于具有很少社交活动信息的新网络[128]和新用户[126]，其他源网络所传递的信息可以在预测链接时极大地克服冷启动问题。更重要的是，跨越对齐网络的共享实体的信息可以为它们所形成的社团结构提供更全面的知识。通过利用跨多个对齐网络的信息，Zhang等人[114]提出一种新模型，用于将共享实体的聚类结果与其他对齐网络中的信息重新组合。Jin等人[37]提出可扩展的框架来研究多个对齐的大规模网络的协同划分，考虑不同网络间的关系，并保持一定的一致性，将不同网络的相同节点划分为相同分区。Zhang等人[122]利用从其他对齐网络传输的信息来研究新兴网络中的社团检测，以克服冷启动。另外，通过融合多个异质社交网络，用户可以通过网络内连接（如，用户间的关系连接）和网络间连接（即，跨网络的映射）彼此广泛相连。因此，信息可以覆盖更多用户，并在社交网络中实现更大的影响。Zhan等人在[119]中提出一种新模型来研究跨多个对齐网络的信息扩散过程，并发现[120]中对齐网络的关键用户。

**2.8 其他应用**

除了上面讨论的任务外，还有其他在异质网络中的应用，如影响传播和隐私风险问题等。为从异质网络中定量地学习影响，Liu等人[53]首先使用生成图模型来学习直接影响，然后使用传播方法来挖掘间接和全局影响。通过使用元路径，Zhan等人[119]提出模型M＆M来解决多部分对齐的异质在线社交网络中影响最大化问题。对于匿名异质信息网络的隐私风险，Zhang等人[121]提出一种去匿名攻击，利用已识别的漏洞来捕获风险。针对无监督文本嵌入方法的性能较差问题，Tang等人[89]提出用于文本数据的半监督表示学习方法，其中标记信息和不同级别的单词共现信息被建模为大规模异质文本网络。为提高线下销售的有效性，Hu等人[25]基于语义的元路径构建公司-公司图，然后在图上采用标签传播来预测有前途的公司。

**北 京 邮 电 大 学**

**本科毕业设计（论文）开题报告**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 学院 | 计算机学院 | 专业 | 计算机科学与技术 | | | 班级 | 2015211307 |
| 学生姓名 | 王睿嘉 | 学号 | 2015211906 | | | 班内序号 | 31 |
| 指导教师姓名 | 石川 | 所在单位 | 计算机学院 | | | 职称 | 教授 |
| 设计（论文）题目 | （中文）异质信息网络中的相似性推荐算法研究 | | | | | | |
| （英文）Research on Similarity Recommendation Algorithm in Heterogeneous Information Network | | | | | | |
| 毕业设计（论文）开题报告内容：  **一、 选题背景和意义**  推荐是解决信息过载的有效方法，被广泛应用于电子商务和互联网服务中。而融合更多信息进行混合推荐是解决数据稀疏性的有效技术。  近年来，大量涌现的社会媒体网站包含许多不同类型对象间的复杂交互，将这些对象建模为同质网络是很困难的。作为半结构化的表示方法，异质信息网络可以有效建模和处理复杂多样的数据。其次，与同质网络相比，异质网络可以融合更多类型的对象及其复杂交互关系，也可以融合多个社交网络平台的信息。最后，异质信息网络包含丰富的语义，将导致更细微的知识发现。因此，异质信息网络作为有效的信息融合方法可以用于整合推荐系统中的对象和关系，可能产生更加准确的推荐结果。  本工作首先调研总结相关异质信息网络推荐算法，了解异质信息网络的数据形式及推荐评价方法。然后，深入挖掘不同元路径下的语义关系相似性，设计基于元路径的推荐算法，使得推荐模型融合多种语义信息，得到更加全面准确结果。   1. **研究现状和目标** 2. **研究现状**   由于大量复杂系统存在广泛的关联数据，信息网络建模与分析成为学术界和工业界的研究热点。很多研究者从不同角度对信息网络展开研究。下面从与本工作密切相关的两个方面，即异质信息网络中的相似性推荐和网络表示学习，分别介绍国内外研究现状。  1.1 异质信息网络中的相似性推荐  传统的推荐系统通常仅利用用户-产品评分反馈信息进行推荐。协同过滤是最普及的技术之一，包括两种方法：基于记忆的方法和基于模型的方法。最近，矩阵分解体现出其在推荐系统中的有效性和效率，它将用户-产品评分矩阵分解为两个低级的用户-特征和产品-特征矩阵，然后利用分解矩阵进行预测。随着社交媒体的普及，越来越多的研究人员通过利用用户间的社交关系研制社交推荐系统。  近年来，异质信息网络（Heterogeneous Information Network），即多种类型节点和链路的建模方式，在数据挖掘领域大放异彩。作为重要的数据挖掘任务，推荐系统包括许多对象类型（如电影推荐中的用户、电影、演员和导演等）及对象间的丰富交互关系（如观看信息、社交信息和属性信息等），可以自然而然地利用异质信息网络建模。而异质信息网络优越的信息综合能力及丰富的语义信息使得更准确的推荐成为可能。  元路径（meta-path），即连接两类对象的关系组合，被广泛用于异质信息网络中的语义信息建模和关系抽取。Shi等人实现了一个基于语义的推荐系统HeteRecom，利用元路径的语义信息来评估电影间的相似性。此外，考虑属性值，如链接上的评分，他们进一步将推荐系统建模为加权异质信息网络，并提出基于语义路径的个性化推荐方法SemRec。为充分利用关系异质性，Yu等人引入基于元路径的隐特征来表示沿不同类型路径的用户和产品间的连接，然后利用贝叶斯排名优化技术在全局和个性化水平上定义推荐模型。同样基于元路径，Burke等人提出在加权混合中包含多个关系的推荐方法。  许多方法利用异质信息网络来融合各种信息。结合不同的上下文，Jamali等人提出基于上下文的矩阵分解模型，考虑每个实体的一般隐因子和依赖于上下文的隐因子。使用用户隐式反馈数据，Yu等人进一步解决全局和个性化实体推荐问题。根据相关的兴趣组，Ren等人提出基于簇的引用推荐框架来预测每个查询在书目网络中的引用。同样，Wu等人结合兴趣组信息，利用图表摘要和基于内容的聚类方法来进行媒体推荐。基于多种异质网络特征，Yang等人使用基于SVMRank的方法将多个特征建模至统一的框架中。利用多种类型的关系，Luo等人提出一种社交协同过滤算法。此外，采用用户和产品的相似性作为正规化，一些工作提出用于推荐的矩阵分解框架。  总的来说，大多数异质信息网络的相似性推荐算法集中在简单的网络结构（如二分或星型模式）和粗糙的语义建模（如元路径）上。但很多实际数据，特别是电子商务数据，所构建的异质信息网络包含很多节点和边类型，没有简单的网络模式。这种无模式异质信息网络带来了很多新的研究问题，如多种类型对象关系的管理及元路径的自动产生等，与此同时也亟需更强大语义建模和分析方法挖掘丰富细致的语义信息。  1.2 网络表示学习  网络表示学习是把网络节点映射到一个低维向量空间，从而用低维向量来表示网络中的节点。网络表示学习使得学习到的节点表示能保持网络中的有效结构及性质等信息，并能够与许多传统机器学习技术实现完美耦合，因此成为近些年数据挖掘研究领域的热点问题。  不同于传统的网络表示学习普遍采用浅层模型，图神经网络采用深度神经网络模型分析图数据，具有更好的学习能力，并有可能融合知识实现推理功能，因此成为近些年的研究热点。图卷积网络作为图神经网络的重要分支被广泛研究，其主要分为Spectral方法和Spatial方法。在Spectral方法中，相关研究对图的拉普拉斯矩阵进行特征值分解模拟离散傅里叶变换，将图信号变换到谱域来构建在图结构数据上的“傅里叶基”，从而在图信号上进行卷积操作。ChebNet模型在图谱论基础上，针对傅里叶基运算耗时问题，设计了基于契比雪夫多项式的滤波器，保证算法的时间复杂度与图规模线性相关。GCN模型将ChebNet进行了一阶近似，将邻接矩阵更换成自环的邻接矩阵；作为ChebNet的一种特殊形式，GCN将谱图卷积和空间结构结合以得到更好的节点表示。在Spatial方法中，相关研究使用图的拓扑结构，将相邻节点看作一个集合，节点之间通过特征信息的相互传递，从而学习每个节点的表示。同时，由于注意力机制在许多任务中取得了效果，研究者也试图在图神经网络中使用注意力机制。  总体来看，目前大多数的网络表示学习重点考虑如何更有效地保持网络的结构以及性质等，并在许多任务上都取得优越的效果。因此，如何将先进的网络表示学习方法更好地运用至异质信息网络中的相似性推荐，将是本工作的关键研究内容。   1. **拟解决的主要问题**   围绕研究目标，学习异质信息网络中的相似性推荐、网络表示学习等相关理论知识，掌握相应算法原理，并尝试在异质信息网络上运用GCN产生更好的推荐结果；  学习并搭建深度学习框架TensorFlow；  编程实现模型，并准备经典数据集对模型进行训练。验证算法的有效性，并对结果进行评估，尝试总结模型优缺点；  对算法的不足之处尽可能进行优化。  **三、 研究方法和措施**  根据本课题的研究内容和拟解决的主要问题，决定采用以下方法进行研究：   1. **文献调研**   异质信息网络中的相似性推荐是未接触过的新问题。首先，阅读相关综述，了解异质信息网络中的重要概念，从而对待研究的问题产生宏观上的总体认知。其次，通过进一步阅读相应论文，学习综述所提及的先进且具有代表性的算法，并根据研究目标认真思索解决方法，尝试对现有模型进行改进。   1. **博客参考**   在了解相似性推荐算法的基本实现方法后，对于论文中阐述不够清晰或模型设置不够明确的部分，通过查找相关视频或博客进一步学习。   1. **源码阅读**   参考GitHub上相似性算法的具体实现，从他人的模型中思考如何实现自身的研究目标，如何在已有的基础上实现改进，在阅读源码过程中寻找灵感。   1. **交流讨论**   在模型设计过程中遇到问题或疑惑，可与导师或师兄交流讨论，在其指导下找到相应的解决方法。   1. **代码实现**   有了详细的设计方案后，可以开始算法代码的编写。在实际操作中进一步熟悉框架的使用方法、参数的优化过程等，并在迭代中观察模型的效果，尝试进一步改进与调整。  **四、 进度安排**  第一阶段（01-02周）：阅读相关论文，了解相似性推荐的实现算法，并提交开题报告；  第二阶段（03-04周）：根据调研结果，对比分析现有的相似性推荐算法，选择合适的算法，并在此基础上改进，进一步设计详细的实现方案；  第三阶段（05-08周）：熟悉数据特点及相应算法框架，完成方案算法的编程实现，进行中期检查；  第四阶段（09-11周）：完成模型实验验证模块的编程实现，并对结果进行评估；  第五阶段（12-15周）：完成毕业设计论文的撰写，进行答辩。  **五、 主要参考文献**  1. Shi C, Li Y, Zhang J, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 17-37.  2. Shi C, Zhou C, Kong X, et al. Heterecom: a semantic-based recommendation system in heterogeneous networks[C]//Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1552-1555.  3. Shi C, Zhang Z, Ji Y, et al. SemRec: a personalized semantic recommendation method based on weighted heterogeneous information networks[J]. World Wide Web, 2019, 22(1): 153-184.  4. 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. ACM, 2014: 283-292.  5. Burke R, Vahedian F, Mobasher B. Hybrid recommendation in heterogeneous networks[C]//International Conference on User Modeling, Adaptation, and Personalization. Springer, Cham, 2014: 49-60.  6. Jamali M, Lakshmanan L. HeteroMF: recommendation in heterogeneous information networks using context dependent factor models[C]//Proceedings of the 22nd international conference on World Wide Web. ACM, 2013: 643-654.  7. Ren X, Liu J, Yu X, et al. Cluscite: Effective citation recommendation by information network-based clustering[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 821-830.  8. Wu J, Chen L, Yu Q, et al. Trust-aware media recommendation in heterogeneous social networks[J]. World Wide Web, 2015, 18(1): 139-157.  9. 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.  10. Shi C, Wang R, Li Y, et al. Ranking-based clustering on general heterogeneous information networks by network projection[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014: 699-708.  11. Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs[J]. arXiv preprint arXiv:1312.6203, 2013.  12. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Advances in neural information processing systems. 2016: 3844-3852.  13. Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.  14. Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017. | | | | | | | |
| 指导教师签字 |  | | | 日期 | 2019 年 3 月 11 日 | | |

**北 京 邮 电 大 学**

**本科毕业设计（论文）中期进展情况检查表**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学院 | | 计算机学院 | 专业 | | 计算机科学与技术 | | 班级 | 2015211307 |
| 学生姓名 | | 王睿嘉 | 学号 | | 2015211906 | | 班内序号 | 31 |
| 指导教师姓名 | | 石川 | 所在单位 | | 计算机学院 | | 职称 | 教授 |
| 设计（论文）题目 | | （中文）异质信息网络中的相似性推荐算法研究 | | | | | | |
| （英文）Research on Similarity Recommendation Algorithm in Heterogeneous Information Network | | | | | | |
| 目前已完成任务 | 1. **调研研究现状，深入理解选题背景和意义**   推荐是解决信息过载的有效方法，被广泛应用于电子商务和互联网服务中。而融合更多信息进行混合推荐是解决数据稀疏性的有效技术。  近年来，异质信息网络（Heterogeneous Information Network），即多种类型节点和链路的建模方式，在数据挖掘领域大放异彩。作为重要的数据挖掘任务，推荐系统包括许多对象类型（如电影推荐中的用户、电影、演员和导演等）及对象间的丰富交互关系（如观看信息、社交信息和属性信息等），可以自然而然地利用异质信息网络建模。而异质信息网络优越的信息综合能力及丰富的语义信息使得更准确的推荐成为可能。  元路径（meta-path），即连接两类对象的关系组合，被广泛用于异质信息网络中的语义信息建模和关系抽取。Shi等人实现了一个基于语义的推荐系统HeteRecom，利用元路径的语义信息来评估电影间的相似性。此外，考虑属性值，如链接上的评分，他们进一步将推荐系统建模为加权异质信息网络，并提出基于语义路径的个性化推荐方法SemRec。为充分利用关系异质性，Yu等人引入基于元路径的隐特征来表示沿不同类型路径的用户和产品间的连接，然后利用贝叶斯排名优化技术在全局和个性化水平上定义推荐模型。同样基于元路径，Burke等人提出在加权混合中包含多个关系的推荐方法。  但是，现有异质信息网络的推荐算法大多聚焦于用户和物品间的交互信息，即购买历史，而这在实际应用中仅能反映用户口味和物品特性的一个方面。因此，为深入挖掘不同方面的语义关系相似性，**Neural network based Aspect-level Collaborative Filtering model (NeuACF)**进一步设计基于元路径的推荐算法，使得推荐模型融合多方面语义信息，得到更加全面准确的结果。  但NeuACF中，用户和物品的交互可能性是利用embedding间的点积来度量的。由于点积本身所固有的缺陷——不满足三角不等式，可能导致相似性的传递特征被破坏，从而限制模型效果的提升。与此同时，NeuACF使用point-wise损失函数，关注于交互可能性的绝对数值。但对于Top-N推荐而言，相对交互可能性更为重要，因而通过修改损失函数显式捕捉排序信息，可能产生更为精准的推荐列表。   1. **确定具体改进方案，并根据实际需求选择数据集，编码实现模型**   模型分为三个模块，即首先抽取aspect-level特征，其次进行表示学习并融合，最后利用适当打分和损失函数完成优化，总体框架如下图所示。接下来，分模块阐述算法原理、与NeuACF的异同及相应实现进度。     1. **基于PathSim的aspect-level相似性矩阵计算（NeuACF）**     给定元路径，需要选择合适的算法抽取aspect-level特征。本工作遵循NeuACF，利用PathSim计算相似性矩阵，并作为特征输入后续模型。原因有二，其一，基于相似性可以缓解噪音；其二，利用相似性矩阵可以抽取高阶特征。  为此，查找PathSim源码，并在MATLAB运行，完成该模块的编程实现。   1. **基于Multi-Layer Perception（MLP）的aspect-level表示学习（NeuACF）**     相对于浅层模型而言，深层模型优越的非线性关系捕捉能力使其应用范围越来越广。因此，利用MLP将相似性矩阵作为输入，并输出相应aspect-level embedding：    其中，是相似性矩阵B中用户i的对应向量，是MLP第l层的隐向量，和是相应第l层的权重和偏置矩阵，是MLP输出的用户embedding。与此同时，所有层的激活函数均选择ReLU。  与NeuACF完整模型不同，本阶段暂时以average方式融合aspect-level embedding，完成模型简化版本的编码实现，方便调试与正确性检验。后续阶段再加入attention机制，考虑不同aspect的重要性，进一步提升模型效果。   1. **引入metric learning和pair-wise损失函数（改进）**   首先，NeuACF利用点积度量用户和物品间的交互可能性：    其中是用户embedding，是物品embedding。因为点积不符合三角不等式的约束，因此可能无法反映更细粒度的用户口味，从而降低模型准确度。以一个简单的例子来说明，如下所示：    表格是用户对物品的偏好信息，左图展示基于点积的矩阵分解方法的一组稳定解，即若用户喜爱物品，则他们间点积为2，否则为0。此时，可以观察到，同时喜欢和，但和间点积为0，即用户对物品的偏好虽然得以体现，但用户和用户、物品和物品间的相似性可能遭到忽视，而这一点对于推荐至关重要。  为此，引入距离的概念。距离，是许多机器学习算法的核心概念，如K-nearest neighbor（kNN）、K-means等。Metric learning技术利用距离定义的metric捕捉数据间的关系，从而使得相似对间距离较近，不相似对间距离较远。而metric所需满足的重要性质之一——三角不等式，自然地蕴含相似性的传递过程，即将已知的相似性信息通过距离传递给未知的点对，这与协同过滤的本质思想不谋而合。因此，**Collaborative Metric Learning（CML）**将metric learning应用至协同过滤，通过用户-物品metric，在反映用户偏好的同时，捕捉用户-用户及物品-物品间的相似性。因此，本工作考虑以metric替换NeuACF中的点积估计形式，实现更全方位的信息传递。具体地，参考CML，修改NeuACF打分函数：    其中是用户embedding，是物品embedding。  其次，NeuACF利用point-wise损失函数进行模型优化：    其中，和分别是正负实例集合，是预测分数，是ground truth。在此种情况下，选取top-ranked结果作为最终的推荐集合。但point-wise损失函数作为预测结果与ground truth的交叉熵，更为关注分数的绝对数值，较适合基于显式反馈的rating推荐系统。而基于隐式反馈的Top-N推荐系统注重物品列表的相对分数。因此，本工作进一步考虑使用基于负采样的pair-wise损失函数替换NeuACF中的point-wise损失函数，产生排序更合理的推荐列表。具体地，在metric基础之上，与不喜欢的物品相比，用户应与喜欢的物品有更小的距离。为此，修改损失函数为：      其中，是正样本集合，是超参数margin。  以上两点即为模型的优化部分。至此为止，编码实现了NeuACF改进模型的简化版本（无attention）。   1. **数据集及编程环境说明**   为验证模型有效性，选择公开的MovieLens和Amazon数据集。其中，前者广泛用于电影推荐系统，后者用于商品推荐。除电影和用户外，NeuACF还需导演和演员信息来学习基于元路径的aspect-level节点表示，而这些信息不包含在原数据集中。因此，利用爬虫工具，从IMDb上抓取相应电影的导演和演员列表补充至MovieLens。  进一步，基于Python 3.6搭建TensorFlow 1.2.1环境，并参考NeuACF源码和官方指导手册，熟悉TensorFlow的函数使用方法。与此同时，实现部分官方样例，在动手操作中，加深对框架的理解和体会。 | | | | | | | |
| 是否符合任务书要求进度 是 | | | | | | | |
| 尚需完成的任务 | 1. 进一步优化模型，实现以attention方式融合aspect-level embedding； 2. 在不同数据集下验证模型有效性，并与baseline（NeuACF）进行比较； 3. 撰写毕业设计论文，进行答辩。 | | | | | | | |
| 能否按期完成设计（论文） 能 | | | | | | | |
| 存在问题和解决办法 | 存  在  问  题 | 1. 对深度学习框架TensorFlow的了解不够深入，模型搭建过程难点颇多； 2. 参数较多，难以确定最佳参数组合。 | | | | | | |
| 拟  采  取  的  办  法 | 1. 根据遇到的实现难点，反复阅读TensorFlow官方指导手册、博客或GitHub中相关源码，理解并熟悉函数接口； 2. 采用控制变量法或网格搜索确定最优参数组合。 | | | | | | |
| 指导教师  签字 | |  | | 日期 | | 年 月 日 | | |
| 检查小组  意见 | | 负责人签字： 年 月 日 | | | | | | |