

杭州电子科技大学

硕士学位论文

题目：基于迁移学习的跨领域
推荐的方法研究

研究生 王 欣

专 业 计 算 机 技 术

指导教师 万 健 教 授

完成日期 2015 年 05 月

杭州电子科技大学硕士学位论文

**基于迁移学习的跨领域
推荐的方法研究**

研 究 生： 王欣

指导教师： 万 健 教授

2015 年 05 月

**Dissertation Submitted to Hangzhou Dianzi University
for the Degree of Master**

Research on Methods of Cross-Domain Recommendation with Transfer Learning

Candidate: Wang Xin

Supervisor: Prof. Wan Jian

May, 2015

杭州电子科技大学

学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。

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。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

本人完全了解杭州电子科技大学关于保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属杭州电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署各单位仍然为杭州电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。

（保密论文在解密后遵守此规定）

论文作者签名： 日期： 年 月 日

指导教师签名： 日期： 年 月 日

摘要

随着计算机技术的发展和网络的不断普及,人们已处在信息大爆炸的时代。推荐系统就是在这种背景下应运而生。推荐系统从众多的信息中为用户推荐用户本身感兴趣的内容,提高了用户获取信息的效率。目前,大多数推荐系统都是基于协同过滤推荐技术来进行推荐,但它存在两个主要的问题,即数据稀疏性问题和冷启动问题。

目前,有一些学者提出了运用迁移学习的思想去解决协同过滤推荐系统中的数据稀疏性问题和冷启动问题。但这些模型几乎都没考虑两个领域间的差异性,如评分刻度差异性、领域间相关性。这种差异性可能会导致数据的负迁移,造成推荐准确度降低。针对这一问题,本文提出了两种新的基于迁移学习的跨领域推荐方法:

1) 提出一种基于标签的跨领域推荐方法。大多数推荐系统都有对用户和物品进行描述的标签,本文利用这些标签来连接两个领域并进行数据的迁移,因为标签记录的是用户特征,所以领域间知识的迁移就是对用户特征进行迁移,因而避免了领域间的评分刻度差异所带来的问题,从而解决了领域间的差异性导致的数据负迁移。该方法主要分为三步:首先利用非负矩阵分解算法对辅助数据域进行用户分类,得到不同类型的用户分组;其次,通过 BP 神经网络对用户的特征进行学习,训练得到根据用户的特征来判断用户类型的神经网络;最后,利用训练好的 BP 神经网络对目标数据域的用户进行类型判断,并根据同类型用户对同一物品的平均评分去填充目标数据域的空缺值。

2) 提出一种基于潜在特征聚类的跨领域推荐方法。该方法不仅学习不同领域间的“共享知识”,而且还学习不同领域间的“特殊知识”,“特殊知识”代表属于每个领域自己的那部分“知识”,并且这部分“特殊知识”可以解决领域间的不相关性所导致的推荐准确度降低这一问题。同时,本文通过 DP 距离匹配算法去计算两个领域之间的相似度,通过该相似度来调整领域间共享维度和私有维度的比例,进一步提高了方法的效率。

关键词: 迁移学习, 跨领域推荐, 标签, BP 神经网络

ABSTRACT

With the development of computer technology and the expansion of the network, people has been surrounded by the sea of information. Recommendation system was advanced under this background. Recommendation system is a hot research field, and it has been successfully applied on E-commerce sites. It recommended contents for users which they are interested in numerous information. To make the users get the information quickly. At present, most of the recommendation system is based on collaborative filtering recommendation technology. But there are exist two main problems: data sparsity problem and cold start problem.

At present, some models based on transfer learning have been proposed to alleviate the data sparsity problem and cold start problem. But they do not take the diversity among the related domains into account. The diversity might clean the advantages of common pattern, which may result in bad performance. For the problem, we proposed two new cross-domain recommendation methods with transfer learning:

1) This thesis proposes a transfer learning method based on tags. Most of recommendation have tags which used to describe the features of users and items. We can exploit the tags to connect the two domains. The algorithm is mainly divided into three steps: Firstly, we classify the users from auxiliary domain by non-negative orthogonal matrix tri-factorization, so we can get the different type of user; Then we train the BP neural network through the feature tags of different type of users, and the BP neural network can judge the type of user by their features. At last, we get all the types of the users from target domain by the trained neural network, then predict the missing value in target matrix by the average value of the rating which the same type of user rating on the same item.

2) This thesis proposes a cluster-level based latent factor method. This method can not only learn the common rating pattern shared across domains with the flexibility in the controlling the optimal level of sharing, but also learn the domain special rating patterns of users in each domain that involve the discriminative information propitious to performance improvement.

Keywords: transfer learning, cross-domain recommendation, tag, bp neural network

目 录

摘要.....	I
ABSTRACT.....	II
第 1 章 绪论.....	1
1.1 研究的背景及意义.....	1
1.2 国内外的研究现状及分析.....	2
1.2.1 迁移学习的研究进展.....	2
1.2.2 跨领域推荐的研究进展.....	3
1.3 论文的主要工作.....	4
1.4 论文的组织与结构.....	4
第 2 章 迁移学习及跨领域推荐相关技术综述.....	6
2.1 推荐系统.....	6
2.1.1 基于内容的推荐系统.....	6
2.1.2 基于协同过滤的推荐系统.....	6
2.1.3 混合式推荐系统.....	8
2.2 迁移学习.....	9
2.2.1 引言.....	9
2.2.2 迁移学习分类.....	9
2.3 跨领域推荐.....	11
2.3.1 跨领域推荐任务.....	11
2.3.2 跨领域推荐技术分类.....	12
2.4 本章小结.....	14
第 3 章 基于标签学习的跨领域推荐方法.....	15
3.1 概述.....	15
3.2 相关技术.....	16
3.2.1 非负正交矩阵分解聚类算法.....	16
3.2.2 BP 神经网络.....	16
3.3 基于标签的跨领域协同过滤推荐算法.....	17
3.3.1 用户（物品）聚类.....	17
3.3.2 基于 BP 神经网络的特征学习.....	20
3.3.3 对目标数据域的用户评分预测.....	22

3.4 实验评估.....	23
3.4.1 实验设计.....	23
3.4.2 实验数据及预处理.....	24
3.4.3 对比模型.....	25
3.4.4 实验评估方法.....	26
3.4.5 实验结果及分析.....	26
3.5 本章小结.....	28
第4章 基于潜在特征聚类的跨领域推荐方法.....	30
4.1 概述.....	30
4.2 问题定义.....	31
4.3 基于潜在特征聚类的跨领域推荐方法.....	31
4.3.1 CBT 模型.....	31
4.3.2 CBT 模型优化.....	32
4.3.3 领域相关性分析.....	33
4.4 实验评估.....	35
4.4.1 实验数据.....	35
4.4.2 对比模型.....	35
4.4.3 实验设计.....	36
4.4.4 实验评估方法.....	36
4.4.5 实验结果及分析.....	37
4.5 本章小结.....	38
第5章 总结与展望.....	39
5.1 工作总结.....	39
5.2 进一步工作及展望.....	40
致 谢.....	41
参考文献.....	43
附录：.....	47

第 1 章 绪论

1.1 研究的背景及意义

随着计算机技术的发展和互联网规模的迅速扩大，出现了越来越多的网络信息，人们在面对爆炸式信息的选择时，往往都不知所措，很难关注到或者找到他们需要了解的那些信息，信息的增多反而导致了人们获取信息效率的下降。而推荐系统就是解决信息超载问题的一个很好的方案。目前，大多数推荐系统都是基于协同过滤推荐技术进行推荐的，协同过滤推荐是根据物品的特征以及用户的偏好来向用户产生推荐的，它通过一些相似度计算方法来得到和用户偏好较为相似的一群用户，并且根据这群用户中的若干用户的喜好程度来推荐物品；或者得到和用户喜爱的物品较为相似的一组物品，选取相似度较高的一些物品进行推荐。推荐系统通过用户的偏好来发现用户喜爱的物品并进行推荐，在一定程度上解决了信息超载的问题。

但是，现在绝大多数的推荐系统都只在单个领域进行推荐，如 Netflix 只推荐电影和电视系列，并且基于协同过滤的推荐系统往往都存在两个问题，即数据稀疏性问题和冷启动问题。其实，协同过滤推荐系统是利用用户对物品的评价来得到用户偏好的模型，然后根据用户的偏好为他推荐他感兴趣的物品。事实上，用户在不同的领域也会显示出同样的偏好，喜欢看恐怖电影的用户一般情况下也喜欢看恐怖类的书籍。用户在一个领域里的偏好往往能够成功迁移到其他相关领域。通过这种迁移，可以很好的解决某些领域内的数据稀疏性问题和冷启动问题。这是跨领域推荐中一个很重要的任务，也是迁移学习在跨领域推荐中的一个重要应用，但这不是跨领域推荐的全部目标。跨领域推荐的第二个任务是能为用户推荐不同领域的个性化物品。比如，一个推荐系统不仅能为用户推荐电影，并且能为用户推荐与电影相关的音乐、书籍、游戏等。

在本文中，着重研究跨领域推荐中的第一种任务，即通过不同领域的用户偏好迁移来解决某些领域内数据稀疏性和冷启动问题，冷启动问题在一定程度上也可以理解为数据稀疏性问题。所以，数据稀疏性问题就是当前推荐系统中最主要的问题。针对这个问题，本文采用迁移学习的思想，通过迁移辅助数据域中较为稠密的信息到较为稀疏的目标数据域中，以解决目标数据域中的数据稀疏性问题，从而提高推荐系统的推荐准确度。但是，不同领域之间的评分具有差异性，这种差异性可能会造成数据的负迁移，也就是通过数据迁移反而会导致推荐结果更加不准确。因此，本文提出了两种基于迁移学习的跨领域推荐

方法，这两个方法都很好的解决了不同领域间的差异性所导致的推荐结果不好的问题。

1.2 国内外的研究现状及分析

1.2.1 迁移学习的研究进展

传统的数据挖掘和机器学习算法是通过统计模型对收集的训练集进行训练，然后对未知的数据进行预测^[1-3]。半监督分类^[4-7]通常会遇到标记的数据太少而不足以去构建一个好的分类器。对不完整的数据集的监督学习和半监督学习已经被研究。比如 Zhu 和 Wu^[8]研究了怎样处理类标签噪点问题。Yang 等人考虑了对未来样本进行额外的训练时的成本敏感学习^[9]。然而，他们中大多数都是假设标记的数据和未标记的数据都具有相同的分布。迁移学习则相反，它允许域之间、任务之间、训练集和测试集的分布情况可以不相同。在现实生活中，也有很多关于迁移学习的例子。比如，学会如何正确的辨识苹果也会对正确辨识梨有一定的帮助，同样的，学习演奏电子琴也可以帮助我们学习钢琴。人可以通过过去学习的知识去解决另一个新的问题，这也是研究迁移学习的意义所在。

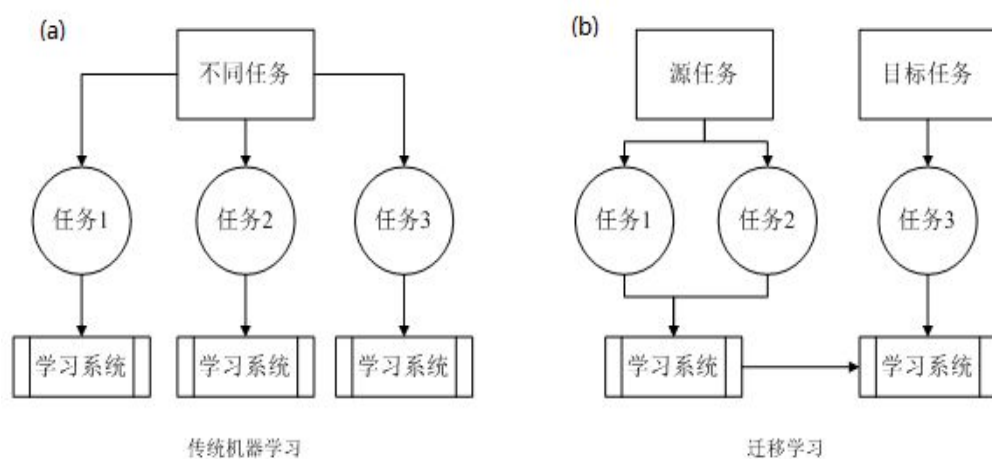


图 1.1 传统机器学习和迁移学习的过程对比

从 1995 年起，就出现了很多技术名词：学会学习、终身学习、知识迁移、归纳学习、多任务学习等。其中多任务学习和迁移学习的相关性最大，一个常见的多任务学习方法就是从中发现潜在特性，然后将这种特性应用到每个任务中去。在 2005 年，BBA 给迁移学习下了一个新的定义，迁移学习是能够把以前学到的知识运用到新的任务中去的能力。在这个定义中，迁移学习的目标就是通过从源任务中抽取出“知识”，然后把抽取的“知识”运用到目标任务中去。和多任务学习不同，迁移学习并不是同时学习源任务和目标任务，在迁移学习

的概念中，源任务和目标任务并不是对称的关系了。图 1.1 展示了传统机器学习和迁移学习的过程的不同。从图中我们可以清楚的看到，传统机器学习是对每个任务都进行学习，而迁移学习则是将之前学到的“知识”运用到另一个任务中去。

1.2.2 跨领域推荐的研究进展

目前，绝大多数推荐系统都只在单个领域进行推荐，如 Netflix 只推荐电影和电视系列。它们利用用户对物品的评价来得到用户喜好的模型。事实上，用户在不同的领域也会显示出同样的偏好，喜欢看恐怖电影的用户大致上也喜欢看恐怖类的书。某个领域里的用户喜好可以成功的迁移到另一个领域，但这只是跨领域推荐的一种任务。跨领域推荐也能为用户推荐不同领域的个性化物品，比如：一个推荐系统不仅能为用户推荐电影，并且能为用户推荐与电影相关的音乐、书籍、游戏等。

跨领域推荐在推荐系统领域中是一个新兴的研究内容，跨领域推荐最早是在 2008 年被 Winoto and Tang 提出来的。在这个工作里^[10]，他们列出了三个基本问题进行调查：1) 验证用户喜好在不同领域的相关性的存在；2) 设计一个能够根据用户在源领域的喜好来预测用户在目标域喜好的模型；3) 制定适当的跨域推荐评估。他们推测，跨领域推荐也许没有单领域推荐精确，但跨领域推荐的结果却更加多元化，这能提高用户对推荐结果的满意度。文中还提到，跨领域推荐技术还有其他的优势，比如解决了冷启动问题^[11-13]和数据稀疏性问题^[14-15]。通过识别物品在不同领域之间的联系，然后利用用户在其他相关领域的偏好为他在一个新的领域提供推荐。事实上，如何利用不同领域的信息为用户进行推荐这个问题在不同的领域有不同的观点，比如用户建模^[16]，信息检索^[10,12,17]，知识管理以及机器学习^[14-15,18,19]等领域。Tobias 等人对跨领域推荐的任务进行了总结^[20]，他们将跨领域任务分为了两种：1) 利用辅助数据域的知识来提高目标数据域的推荐准确度；2) 为用户推荐不同领域中相关的物品。Li 总结了跨领域推荐中的协同过滤方法^[21]，文中区分了三种类型的域：系统域、数据域、时间域。系统域是在构建推荐系统时的那些不同数据集；数据域是用户偏好的不同表现形式，它可以是隐式的（例如点击、购买）或显性的（例如确定的评分值）；时间域就是按时间划分的那些子集。

目前，迁移学习已经被用来解决跨领域推荐中第一种任务，也就是解决目标域中数据稀疏性问题。现在已经有很多跨领域推荐模型被提出，这些模型主要通过从辅助领域中抽取“知识”，并将抽取的“知识”迁移到目标领域中，从而解决目标域中数据稀疏性问题。

1.3 论文的主要工作

针对推荐系统中的数据稀疏性问题，本文提出了两个基于迁移学习的跨领域推荐方法，并且这些方法解决了当前跨领域推荐模型中未考虑领域间差异性所导致的数据负迁移问题。本文的主要工作包括下面两个内容：

1) 提出一种基于标签学习的跨领域推荐方法

该方法是利用共享的标签来连接两个不同领域，通过迁移辅助领域的用户偏好来减轻目标域的数据稀疏问题，从而提高推荐质量。在该方法中，考虑了不同领域之间的差异性。并且在公有的推荐系统数据集上验证了该方法能通过迁移用户偏好解决数据稀疏性问题，并且在评估指标中优于其他的单领域推荐模型和跨领域推荐模型。同时，本文还验证了辅助域的数据稀疏性对迁移效果的影响，即辅助域的数据稀疏性越高，算法的迁移效果越不好。

2) 提出一种基于潜在特征聚类的跨领域推荐方法

该方法的核心思想是从一个辅助数据域中获取“知识”，然后将该知识迁移到目标数据域中，从而解决目标数据域中的数据稀疏性问题。在该方法中，将获取的“知识”分为“共享知识”和“私有知识”，通过每个领域的“私有知识”来消除领域间的差异性，解决领域间的不相关性所带来推荐准确度下降等问题，并且通过 DP 距离匹配算法计算两个领域的相似度，根据该相似度来确定共享维度和私有维度，提升了方法的效率。在实验中验证了该方法能够解决推荐系统中数据稀疏性这个问题。

1.4 论文的组织与结构

本文在第一章首先介绍了网络上信息高速增长带来的信息超载问题，从而引出推荐系统，并指出了当今基于协同过滤推荐系统的不足，数据稀疏性问题和冷启动问题。接着介绍了迁移学习和跨领域推荐的研究进展，并重点描述了运用迁移学习的思想来解决跨领域推荐中的数据稀疏性问题。最后，针对跨领域推荐模型中的不足（未考虑领域间的差异性所带来的数据负迁移）提出了两个基于迁移学习的跨领域推荐方法。

本文在余下的内容主要分为四个部分：

第二章首先主要介绍了推荐系统的历史和算法，然后讨论了迁移学习的分类和迁移学习研究的三个主要内容：what to transfer、how to transfer、when to transfer。最后详细描述了跨领域推荐的两种任务以及跨领域推荐的技术分类。

第三章主要介绍了本文提出的基于标签的跨领域推荐方法，解释了利用标

签是如何避免领域间评分刻度的差异性所带来的数据负迁移问题。并详细介绍了该方法的三个部分。首先，根据辅助数据域较为稠密的评分信息对用户进行分类；其次，对辅助数据域中已分类好的用户进行特征学习(利用标签信息)；最后，利用训练好的神经网络以及目标域中的用户标签对用户进行归类，将目标数据域中相同类型的用户评分平均值填充同类型用户的评分空缺值；随后对该算法在公开数据集上进行了实验，并对实验结果在迁移效果上和辅助域的数据稀疏性对迁移的影响两个方面进行了分析。

第四章描述了一个基于潜在特征聚类的跨领域推荐方法。该方法主要是通过从辅助域中获取“知识”，并将获取到的“知识”迁移到目标域中，最重要的是在整个过程中，考虑了不同领域之间的差异性。最后通过实验对迁移效果进行了分析。

第五章是对本文进行了总结，分析了文章的贡献并指出了其中的不足。并且在本文的基础上对进一步的工作进行了展望。

第2章 迁移学习及跨领域推荐相关技术综述

2.1 推荐系统

计算机技术的发展和网络规模的迅速扩大将人们带入了信息大爆炸的时代，人们面对铺天盖地的选择时往往却不知道如何选择。而推荐系统的出现，解决了人们获取信息难的问题。它通过对用户的偏好和物品的特征进行分析来得到用户感兴趣的一些物品，并将其中最感兴趣的物品推荐给用户。推荐系统是一个很热门的研究领域，并且成功的运用于电子商务网站，如 Amazon、淘宝、天猫等。目前，推荐系统中主要的推荐方法包括：基于内容的推荐、基于协同过滤技术的推荐以及混合式推荐。

2.1.1 基于内容的推荐系统

基于内容的推荐是根据内容的信息来进行推荐的，它根据内容并且运用机器学习等算法挖掘出符合用户兴趣爱好的物品，并且它不需要用户对其他物品的评价等信息。例如，在书籍推荐中，系统首先会分析用户感兴趣的书籍所具有的特征（作者、风格），然后根据书籍特征去寻找与之相似的其他书籍，最后向用户进行推荐。基于内容的推荐算法包括：基于关键词的空间向量模型、Rocchio 算法等。

基于内容推荐方法的优点是：

- 1) 不存在冷启动问题和数据稀疏性问题。
- 2) 推荐的物品很符合用户的需求，不随大流。
- 3) 可以向用户推荐新的物品或者不是热门的物品。
- 4) 可以很清楚的了解到被推荐物品的意义。
- 5) 基于内容的推荐技术相对其他技术而言较为完善。

基于内容的推荐的缺点是内容需要有一定的结构性，系统能够从中抽取有代表性的特征，并且要求用户的偏好也具备这种特性，而且仅限为当前用户进行推荐，不能得到其他用户的偏好等。

2.1.2 基于协同过滤的推荐系统

协同过滤技术是目前大多数推荐系统中较为常用的一种推荐技术，例如 amazon.com、taobao.com、tmall.com 等。协同过滤推荐依靠用户对一系列物品的历史评分记录去为用户推荐物品，它通过一些相似度计算方法来得到和用户偏好较为相似的一群用户，并且根据这群用户中的若干用户的喜好程度来推荐

物品；或者得到和用户喜爱的物品较为相似的一组物品，选取相似度较高的一些物品进行推荐。整体上，可以将协同过滤算法划分为 Memory-Based 协同过滤算法和 Model-Based 协同过滤算法。分类详情见图 2.1。

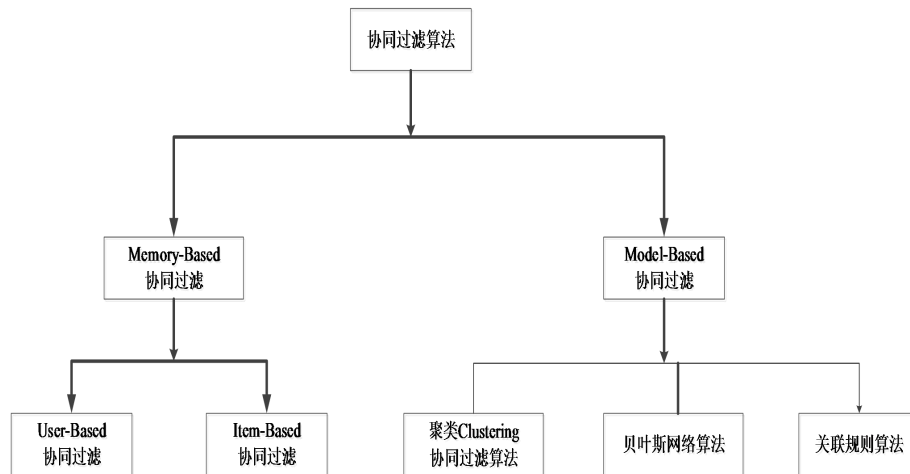


图 2.1 协同过滤算法的分类

● Memory-Based 的协同过滤推荐算法

Memory-Based 协同过滤算法主要分为 User-Based 协同过滤和 Item-Based 协同过滤，下面详细介绍这两种协同过滤推荐算法：

1) User-Based 的协同过滤推荐算法

User-Based 协同过滤算法^[22]通过相似度计算公式得到用户之间的相似度，然后根据相似度得到与用户兴趣爱好较为相似的一群用户，并且从中选取一些用户，如 TOP-N 算法，根据选取的用户的兴趣爱好为当前用户推荐物品。

2) Item-Based 的协同过滤推荐算法

随着时间的推移，推荐系统中的用户数量会越来越多，这就导致了计算相似度的时间会越来越长，而推荐系统中的物品数量则相对用户而言较为稳定。因此，Item-Based 的协同过滤算法就被 Sarwar^[23]提出来了。在基于物品的协同过滤算法中，通过计算物品之间的相似度来得到用户可能感兴趣的其他相似物品，并且将这些物品推荐给用户。例如，喜欢 iphone 的用户一般都喜欢 ipad，因为它们具有相同的特性，例如都是 IOS 系统。

● Model-Based 的协同过滤推荐算法

Memory-Based 协同过滤算法有一个缺点，那就是当数据较少时，在计算相似度过程中往往不准确，这也是目前大多数基于协同过滤推荐技术的推荐系统存在的问题。为了应对这一问题，Model-Based 协同过滤算法被提出来了。

Model-Based 协同过滤算法是根据已有的数据来得到一个模型，并且通过这个模型进行物品评分的评估。该类型的算法主要包括以下几个算法，如聚类协同过滤算法、贝叶斯网络算法、关联规则算法等，这些算法都是根据已有的

数据得到模型然后进行预测的模型。

表 2.1 推荐方法对比

推荐方法	优点	缺点
基于内容的推荐	推荐结果直观，并且容易理解；	稀疏性问题；冷启动问题 只能处理简单属性； 需要较多的分类器
基于协同过滤技术的推荐	能够发现用户的新的感兴趣物品； 数据增多后推荐精度提高； 推荐个性化、自动化程度高； 能处理较复杂的结构	稀疏性问题； 冷启动问题； 可扩展性问题； 质量取决于历史数据集；

2.1.3 混合式推荐系统

基于内容、协同过滤技术的推荐方法都有各自的优缺点，具体情况如表 2.1 所示。因此，在实际情况下，混合式推荐（Hybrid Recommendation）经常被采用。顾名思义，混合式推荐就是将各种推荐方法混合起来进行推荐，以达到提高推荐精度的作用。例如，可以将基于内容的推荐和基于协同过滤技术的推荐结合起来一起使用。混合式推荐的基本思想就是通过不同推荐方法的搭配来弥补各自推荐方法的缺点，起到每种推荐算法扬长避短的作用。但是，混合推荐也并不一定能达到 $1+1>2$ 的效果。目前，混合式推荐技术主要分为下面七种方案：

- 1) 加权（Weight）：加权多种推荐方法结果。
- 2) 变换（Switch）：在实际的推荐过程中根据问题的实际情况变换采用不同的推荐方法。
- 3) 混合（Mixed）：将不同的推荐方法混合起来为用户推荐。
- 4) 特征组合（Feature combination）：通过对不同目标数据源的特征进行组合，以满足某种推荐方法。
- 5) 层叠（Cascade）：通过不同的推荐方法产生的不同推荐结果进行轮番迭代，以达到更高的推荐性能。
- 6) 特征扩充（Feature augmentation）：通过某种技术得到用户或物品的其他特性，从而为推荐方法提供更多的特征输入。
- 7) 元级别（Meta-level）：通过某一种推荐技术得到推荐结果，并将该结果设置为另一种推荐方法的输入。

2.2 迁移学习

2.2.1 引言

数据挖掘和机器学习已经在知识工程领域已经取得了很大的成功，比如分类、回归、聚类^[24-25]。但是，很多机器学习方法都是以一个假设为基础，即训练集和测试集来自相同特征空间和相同的分布。当分布变化时，大多数统计模型需要使用新收集的训练集重新开始重建。在现实的应用中，重新收集训练集并且重建模型花费太大，并且也不大可能。在这种情况下，知识迁移或者迁移学习将是很好的选择。有很多例子可以证实迁移学习在知识工程领域有很大的用处，比如 Web 文档分类^[26-29]。对于一个新的 web 站点进行分类任务时，将会遇到缺乏训练集的问题，因此无法运用一个网页分类器对该站点内容进行分类。对于这种情况，通过迁移知识到该站点将会非常有用。

2.2.2 迁移学习分类

表 2.2 迁移学习的分类及应用场景

迁移学习种类	相关领域	源域数据	目标域数据	任务
归纳式迁移学习	多任务学习	有标记	有标记	回归，分类
	自学习	无标记	有标记	回归，分类
直推式迁移学习	文本分类，样本选择偏置，协方差平滑	有标记	无标记	回归，分类
无监督迁移学习		无标记	无标记	聚类，降维

在迁移学习过程中，主要有三个问题需要解决：

1) 迁移什么 (What to transfer)：What to transfer 是指哪一部分知识可以在不同领域之间被迁移，不同领域之间可能存在一部分共享的知识，而这部分知识可以用来迁移到某一目标领域中去解决目标领域的数据稀疏性问题，从而提高目标推荐系统的推荐准确度。

2) 怎么迁移 (How to transfer)：How to transfer 是指要怎样将那部分共享的知识迁移到另一个领域中去。

3) 什么时候迁移 (When to transfer)：When to transfer 是指在什么情况下能够进行知识迁移，什么情况下不能进行知识迁移。因为当源域和目标域不相关时，盲目的迁移并不能提高目标域的性能，并且还可能会出现负迁移(Negative Transfer)的情况，负迁移会使推荐系统的性能下降。目前大部分迁移学习的研究者都主要在研究“迁移什么”和“怎么迁移”这两个问题，并且它们都有一个共同假设，就是辅助域和目标域之间是相关的。因此，怎样避免负迁移也是当今

一个需要研究的内容。

在迁移学习领域，迁移学习主要被分成了三种，即归纳式迁移学习、直推式迁移学习和无监督迁移学习，具体分类和应用场景如表 2.2 所示。下面，对这三种迁移学习进行详细描述：

1) 归纳迁移学习

在归纳迁移学习中，目标任务不同于源任务，即使源域和目标域是相同的。对于这种情形，需要目标域中有标记的数据去推导得到一个预测模型。此外，可以根据源域中有标记数据和无标记数据将归纳式迁移学习分成两种情况：

- 如果源域中没有或有较少的标记的数据，则归纳式迁移学习可等同于自学习（self-taught learning）^[30]。在自学习中，源和目标域之间的标记空间域可能会有所不同，这意味着源域中的信息不能运用到目标域中，这和源域中没有标记的数据情况类似。
- 如果源域中有较多的标记数据，则归纳式迁移学习和多任务学习较为相似。但是，归纳式迁移学习是通过从源任务中迁移知识来提高目标任务的性能，而多任务学习是同时学习目标任务和源任务。

2) 直推迁移学习

在直推迁移学习中，源任务和目标任务是相同的，但源域和目标域是不同的。在该情况下，目标域没有或者很少有标记过的数据，而源域则有较多的标记数据。可以根据源域和目标域的不同将直推迁移学习也分成两种类型：

- 源域和目标域的特征空间不同。
- 源域和目标域的特征空间相同，但输入数据的边际概率分布是不同的。直推迁移学习的第二种情形和域适应相似，比如文本分类^[31]，样本选择偏置^[32]或者协方差平滑^[33]。

3) 无监督迁移学习

无监督迁移学习和归纳迁移学习相似，目标任务和源任务不同，但和源任务有关。但是，无监督迁移学习主要关注目标域中的无监督学习任务，比如聚类、降维和密度估计^[34-35]。

根据“**What to transfer**”，可以将迁移学习中的迁移方法分为四种，具体分类结果见表 2.3 所示。

1) 基于特征表示的迁移学习

基于特征表示的迁移学习^[44-46]是在目标域中找到一个“好”的特征表示，在这种情况下，不同领域之间迁移的知识编码成学习到的特征表示。通过这些新的特征表示，目标任务的性能将会显著的提高。

2) 基于实例的迁移学习

基于实例的迁移学习^[29,32,36-43]假设源域中的一部分确定数据通过权值的重新分配再学习后，能够重新在目标域中得到利用。因此在这种情况下，存在两个重要的技术，即权重的重新分配与样本的抽选。

3) 基于参数的迁移学习

基于参数的迁移学习^[47-49]是假设源任务和目标任务中共享了一部分参数或支持向量模型的先验分布，并且需要将迁移的知识编码成共享的参数。

4) 基于相关知识的迁移学习

基于相关知识的迁移学习算法^[50]主要是用于相关领域的迁移问题。这个算法的最基本假设是源域和目标域中的数据之间的关系是相似的。因此，迁移知识也就是迁移数据之间的关系。例如，统计关系学习技术^[51-52]。

表 2.3 迁移学习的方法分类

迁移学习方法	详细描述
基于特征表示的迁移学习	找到一个“好”的特征表示，减少源域与目标域的差异，减少分类和回归模型的错误。
基于实例的迁移学习	对辅助域中标记的部分数据通过权重的重新分配应用到目标域中。
基于参数的迁移学习	找到源域和目标域之间的共享参数，这些参数对迁移学习有益。
基于相关知识的迁移学习	在辅助域和目标域之间的建立相关知识的映射。并且这两个域是相关联的。

2.3 跨领域推荐

跨领域推荐在推荐系统领域是一个新兴的研究课题，跨领域推荐最早是 Winoto 和 Tang 在^[10]中提出来的。在那个工作中，他们提出了三个问题，并对这三个问题进行了调查：1) 验证用户在不同领域内的喜好是否具有相关性；2) 设计一个根据用户在源领域的喜好来预测用户在目标域喜好的模型；3) 制定合适的跨域推荐评估。跨领域的推荐可能没有单领域推荐的精度高，但前者更加多元化，这可能会导致更高的用户满意度。更重要的是，跨领域推荐技术还有其他的优点，比如解决冷启动问题和数据稀疏性问题。

2.3.1 跨领域推荐任务

迄今为止，大部分跨领域推荐方法都是基于协同过滤（Collaborative filtering）推荐技术，协同过滤策略利用用户的偏好（通常表示为明确评级项目），而忽略任何基于内容的描述（属性）的物品。当物品属于不同的数据源时，协

同过滤有着巨大的优势。

假设存在两个领域 A 和 B, U_a , U_b 分别表示这两个领域内的用户, I_a , I_b 分别表示这两个领域内的物品, 可以将跨领域推荐任务分成下面两种:

- 1) 利用源域 A 中用户或物品的“相关知识”来提高目标域 B 中的物品推荐的质量。
- 2) 为用户在不同领域内做出联合推荐。例如: 为用户 $U_a \cup U_b$ 推荐物品 $I_a \cup I_b$ 。

2.3.2 跨领域推荐技术分类

Loizou 描述了跨领域推荐主要有三种类型的策略^[53]。1) 利用不同推荐系统中的用户偏好去为其他推荐系统中的用户做出推荐; 2) 根据用户在不同推荐系统的行为, 为他生成多个域联合推荐; 3) 对不同领域的推荐系统进行整合, 建立一个综合的推荐系统。

自从 Loizou 提出自己的观点之后, 有很多跨领域推荐模型被提出来。有些方法使用社交标签和语义知识建立不同领域间用户偏好和物品属性的关系。在不同领域之间没有用户或物品的显式重叠域的情况下, 有些方法通过迁移学习来执行协同过滤推荐。Li 提出一种基于知识迁移并建立不同领域之间的关系的分类^[21], 即集体评级模式、共同的潜在因素、相关性用户/项目的潜在因素。

基于 Li 的分类, Gnacio 对现有的跨领域推荐方法提出了一些新的分类, 他们考虑了跨领域推荐任务定义和域关系类型, 具体分类如下:

- 1) 整合用户在不同领域的偏好

跨领域推荐的一个直接方法是利用用户模型去获取用户在每个领域中对物品的喜好。对于这种情况, 最主要的困难是如何把多个系统中整合到单领域的用户模型。一些作者通过调查策略来获取用户在不同领域的偏好, 然后把它整合到一个单领域的用户模型中。例如, Gonzalez 等人定义了一个特定领域智能用户模型^[54], 它是由一个权重图组成, 描述了不同领域的特征结构, 图中边的权重表示关系的强度。

社会标签系统是一种多域用户偏好的重要来源, Szomszor 等人^[55]提出了一个方法用于过滤和整合标签, 标签与维基百科概念一起创建一个语义模型, 反映用户在多个领域的偏好。

在这一方法类型中, 有一个问题需要进一步的调查, 即整合用户偏好是否真的提高推荐质量, Winton 和 Tang 进行了一个研究^[10], 用户对来自 12 个领域的物品进行评分, 将调查结果汇总后为用户进行联合推荐。得到的结果表明, 从多个领域汇总的信息虽然在推荐准确度上有所下降, 但推荐结果却更加的多样, 并且还有利于推荐系统中的冷启动问题。

2) 利用域间明确的关系

另一种技术是整合多领域的用户偏好，其中包括建立域之间的特征关系，并利用它们来完成跨领域推荐任务。它们中的大多数是基于内容域的关系。Azak^[56]提出的推荐模型，它是利用基于知识的决策规则来建立域之间的关系。图是一个表示域之间的关系的通用结构。Loizou^[53]使用维基百科作为一个通用的词汇来描述物品在多个领域常见的形式。用户和物品被添加到图中的节点上，图中的边代表用户评分和语义项关系。基于这个图，马尔科夫概率模型通过用户和项目之间存在的路径生成推荐。Cremonesi 等人^[17]将属于不同领域的物品置于一个图中，图的每条边连接着两个最相似的物品。因此，为了建立域间关系，某些用户不得不在好几个不同领域对物品进行评分。一般情况下，很少的物品满足这样的条件，这可能会减弱域之间的关系。它们从图的邻接矩阵找到物品之间的非直接关系，并利用它们进行协同过滤推荐。

社会标签也被用来连接领域，因为他们可以作为一个词汇以一个简单、通用的方式来描述任何一个领域的物品。Shi 等人^[57]利用标签来构建用户-用户和物品-物品的相似度矩阵。不同领域的用户/物品之间的相似度和它们之间所共享的标签成正比，计算相似性被合并到一个基于矩阵分解和协同过滤的概率模型中。

3) 跨领域的迁移学习

迁移学习在机器学习领域是一个很热门的研究课题，它通过从一个相关领域迁移知识来提高特定领域的学习任务，在推荐系统领域，迁移学习已经被用来解决协同过滤中的一些问题。在跨领域推荐中，迁移学习主要用来解决协同过滤推荐系统中的数据稀疏性问题，比如 Singh 和 Gordon^[58]，Li, Yang 和 Xue^[14,19]。这些方法主要通过抽取辅助领域的数据，将知识迁移到目标领域中。Phuong 提出将多任务学习运用到协同过滤技术中^[59]，但他们没有考虑辅助领域的来源，他们在同一个协同过滤矩阵中制定了一个多个二元分类的问题，每一个二元分类都对应一个用户。Singh 和 Gordon 在^[58]中提出利用矩阵分解来得到不同领域间通用的潜在特征，并且通过这些特征进行目标数据域稀疏处的填充。Li 等人在^[14]中提出通过一个聚类的 user/item 评分矩阵来充当辅助数据域和目标数据域的桥梁，他们设计了一种聚类压缩算法将辅助数据域的评分矩阵压缩成一个聚类的矩阵 (codebook)，然后，又通过一种新的算法将这个聚类矩阵扩充后填补到目标数据域的稀疏空缺但他们都没考虑辅助领域和目标领域的相关性，不相关的领域进行盲目的迁移会造成负迁移的效果，使推荐精度更低，甚至推荐结果毫处。不相关。Weike Pan 在^[15]中提出了一个基于矩阵分解的框架技术，他们称为坐标系统迁移(CST)，他们设计了 CST 算法，这个算法考虑了

不同数据域中数据的多样性，大体思想还是基于矩阵分解。

从跨领域推荐的研究现状可以看到，当前大部分学者都着重研究跨领域推荐的第一种任务，即通过从辅助域中获取“知识”，然后迁移到目标域中用于解决目标域中的数据稀疏性问题，以提高推荐准确度。大体思路还是从迁移学习的三个最基本问题入手，即 *what to transfer*、*how to transfer*、*when to transfer*，但绝大部分跨领域推荐的研究都集中在前两个问题，很少有跨领域推荐模型中考虑 *when to transfer*。

2.4 本章小结

本章首先介绍了推荐系统出现的背景以及意义，并详细描述了当前推荐系统中的主流推荐算法。其次讲述了迁移学习的概念和迁移学习中需要解决的问题，并对迁移学习的分类、迁移学习中的迁移方法进行了阐述。最后本章分析了跨领域推荐技术的优缺点，总结了跨领域推荐中的任务，并对跨领域推荐的方法进行了分类。

第3章 基于标签学习的跨领域推荐方法

3.1 概述

随着计算机技术的发展和网络规模的不断夸大，人们已被越来越多的信息所包围，而推荐系统的出现解决了获取知识难的问题。但当前大多数推荐系统都是基于协同过滤推荐技术并且是在单一领域为用户进行推荐。协同过滤依靠分析用户的兴趣，找到和自己行为相似的用户，分析这些相似用户的喜好去预测他对其他物品的喜好程度。然而，这一技术最大的困扰就是数据稀疏性问题。因为在实际情况下，大多数用户并不可能对大部分商品进行评分，他们仅仅对其中很小一部分进行评分，新系统尤其明显。这样，通过协同过滤技术进行处理时就很容易导致过度拟合问题，从而造成推荐效果不好。虽然目前有一些基于矩阵分解的技术能够较好的处理数据稀疏性问题，但数据稀疏性问题依旧没能从本质上得到解决。

事实上，用户在不同的领域也会显示出同样的偏好。比如喜欢看恐怖电影的用户大致上也喜欢看恐怖类的书。用户的喜好在不同领域之间的相关性也被统计数据所证实^[10]。所以本文认为不同领域之间共享了相同的评分信息，用户在一个领域里偏好能够成功迁移到其他相关领域。值得注意的是，现在很多推荐系统都会有对用户或物品进行描述的标签，这些标签用于标记该用户的浏览行为或者是用户特性。同样，对物品也会有标签。可以利用标签来作为两个领域的连接点。这样，就避免了两个领域之间评分信息的差异性。本文提出的方法是以两个不同的领域中公用一套相同或相似的标签信息去描述各个领域中的用户特征或者物品特征作为基础。比如，在 `movies` 领域，用来描述电影特征的标签有悲伤、幽默、悬疑、恐怖等。在 `book` 领域，用来描述书的特征的标签也可能包括悲伤、幽默、悬疑、恐怖等。这样，可以利用这些不同领域之间共享的标签来为解决数据稀疏这个问题。

在本章中，提出一种基于标签的跨领域推荐方法，该方法是利用共享的标签来连接两个不同领域，因为标签只记录了用户特征，所以领域间知识的迁移就是对用户特征进行迁移，因而避免了领域间的评分刻度差异所带来的影响，从而提高推荐准确度。该方法具体分为三个部分。

- 1) 根据辅助数据域较为稠密的评分信息对用户（物品）进行分类，因为相同类型的用户对相同的物品的评分大致一样。

- 2) 对辅助数据域中已分类好的用户（物品）进行特征学习。
- 3) 利用训练好的神经网络以及目标域中的用户（物品）标签对用户进行归类，将目标数据域中相同类型的用户评分平均值填充同类型用户的评分空缺值。

3.2 相关技术

本文提出的算法是基于非负正交矩阵分解聚类算法和 BP 神经网络模型。因此，在本节中，将简单介绍下非负正交矩阵分解聚类算法和 BP 神经网络背景知识。

3.2.1 非负正交矩阵分解聚类算法

非负正交矩阵分解算法是建立在非负矩阵分解的基础上，非负矩阵分解算法将矩阵分解成 2 个矩阵，如 $X = FG^T$ ，加上正交的约束后就有更加严谨的聚类解释。但是，这种双正交性有很大的限制。因此需要增加一个额外的矩阵 S 去抵消 X , F , G 之间的差异性。

在非负正交矩阵分解聚类模型中，数据评分矩阵 $X \in R^{M \times N}$ 可以被分解成三个非负矩阵 $U \in R^{M \times K}$ ， $S \in R^{K \times L}$ ， $V \in R^{N \times L}$ ，如 $X \approx USV^T$ 。X 的最优求解过程如下矩阵范式：

$$\begin{aligned} \min_{U, S, V \geq 0} \mathfrak{F}_{ONMTF} &= \|X - USV^T\| \\ \text{s.t. } U^T U &= I, \quad V^T V = I \end{aligned} \quad (3.1)$$

其中 $\|\bullet\|$ 表示 Frobenious 矩阵范式。U, V 都是非负正交矩阵。 $X = [x_1, \dots, x_N]$ 是一个 $M \times N$ 的评分矩阵，它包含了 M 个用户和 N 个物品。从聚类的视角，可以对这三个非负矩阵做如下的理解：

1) $U = [u_1, \dots, u_K]$ 可以表示为潜在用户的特征，其中 u_k 是一个 $M \times 1$ 的向量，表示在 M 个用户中的概率分布，称为 user-cluster 潜在特征。并且 $\arg \max_k (U)_{ik} = k^*$ 意味着第 i 个用户属于 k^* 用户类型。

2) $V = [v_1, \dots, v_L]$ 可以表示为潜在物品的特征，其中 v_l 是一个 $N \times 1$ 的向量，表示在 N 个物品中的概率分布，称为 item-cluster 潜在特征。并且 $\arg \max_l (V)_{il} = l^*$ 意味着第 i 个物品属于 l^* 物品类型。

3) $S = [s_1, \dots, s_K]$ 可以表示为评分矩阵 X 的特征矩阵。

非负正交矩阵分解算法可以同时用户对用户和物品进行聚类，利用用户和物品之间的相互关系完成一个更好的效果。

3.2.2 BP 神经网络

BP (Back Propagation) 网络是一种按误差反向传播的多层前馈网络。它是

由信息的正向传播和逆向传播组成。它的思想就是通过误差的反向传播来不断的动态的调整各层的权值，最终能够得到满足我们期望的输入输出模型。BP 利用一种称为激励函数来描述层与层输出之间的关系，常用的激励函数有 S 型激励函数。BP 神经网络的拓扑结构主要包括三层：输入层、隐层、输出层。下面简单介绍下 BP 神经网络的输入输出模型、作用函数模型、误差计算模型和自学习模型：

1) 节点输出模型

隐节点输出模型： $O_j = f(\sum W_{ij} \times X_i - \theta_j)$

输出节点输出模型： $Y_k = f(\sum T_{jk} \times O_j - \theta_k)$

2) 作用函数模型

作用函数又称刺激函数，一般取[0, 1]内连续取值，如 Sigmoid 函数：

$$f(x) = 1 / (1 + e^{-x})$$

3) 误差计算模型

误差计算模型是反映神经网络期望输出与计算输出之间误差大小的函数： $E_p = 1/2 \times \sum (t_{pi} - o_{pi})^2$

其中， t_{pi} 表示 i 节点的期望输出值； o_{pi} 表示 i 节点计算输出值

4) 自学习模型

神经网络的学习过程，即连接下层节点和上层节点之间的权重矩阵 W_{ij} 的设定和误差修正过程。BP 网络有师学习方式-需要设定期望值和无师学习方式-只需输入模式之分。自学习模型为：

$$\Delta W_{ij}(n+1) = \eta \times \Phi_i \times O_j + a \times \Delta W_{ij}(n)$$

其中， η 表示学习因子； Φ_i 表示输出节点 i 的计算误差； O_j 表示输出节点 j 的计算输出；a 表示动量因子。

3.3 基于标签的跨领域协同过滤推荐算法

3.3.1 用户（物品）聚类

通常情况下，相同类型的用户会对相同类型的物品的评分大致相同。比如一群喜欢看喜剧电影的人会对喜剧类型的电影给出较高的评分，而悲剧类型的电影评分则相对较低。利用这种特性能分类出不同类型的用户。同样的，相同类型的物品被相同用户评的分数也大体一致。在本文中，聚类算法是基于非负正交矩阵分解算法^[60]，在背景中已经介绍过该算法不仅能对用户进行聚类，同时也能对物品进行聚类。本文在此算法基础上稍微修改，使之更好的适用于本算法。

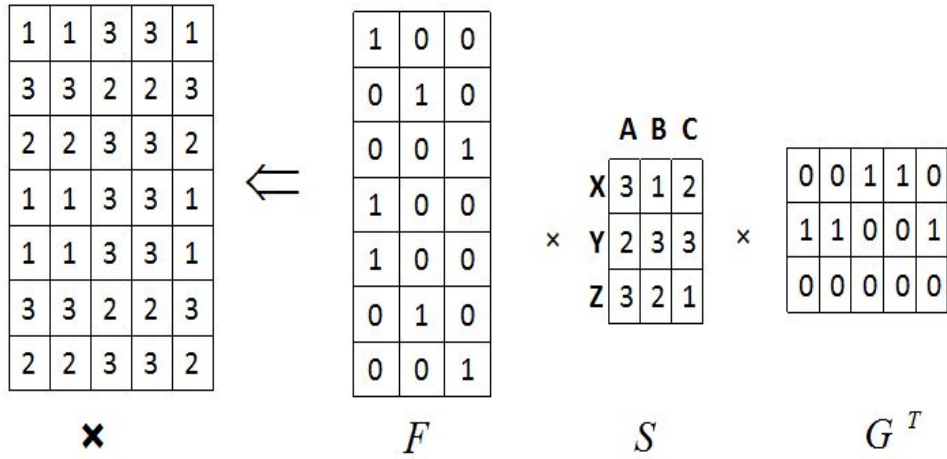


图 3.1 基于 ONMTF 的用户（物品）聚类示意图

辅助数据域的评分矩阵可以被分解为三个非负矩阵相乘。如图 1 所示， X 是一个 $m \times n$ 辅助评分矩阵，它的稀疏度较低。 S 是一个 $k \times l$ 的评分矩阵，其中 k 和 l 不能设置太大或者太小，太大的话矩阵运算时间过大，可能造成不收敛的情况，太小的话可能造成过度拟合的情况。 S 可以理解为 X 的 cluster-level 特征矩阵，其中 X , Y , Z 分别代表三种类型的用户， A , B , C 分别代表三种类型的物品。 F 是一个 $m \times k$ 非负正交评分矩阵，其中行向量代表每个用户的特征信息，相同或相近的行向量可以表示同一类型的用户。如图 3.1 所示， F 中第[1, 4, 5]、[2, 6]、[3, 7]有相同的行向量，它们分别表示三类不同的用户，可以看到相同类型的用户的评分（如矩阵 X 中所示）也是相同的，当然这只是一个最理想的情况。 G 也是一个 $n \times k$ 的非负正交评分矩阵，其中列向量代表每个物品的特征信息，相同的列向量代表同类型的物品，和 F 一样。 X 的近似的求解可以用下列公式优化：

$$\min_{U \geq 0, V \geq 0, S \geq 0} \|X - FSG^T\|_F^2 \quad (3.2)$$

$$G_{jk} \leftarrow G_{jk} \sqrt{\frac{(X^T FS)_{jk}}{(GG^T X^T FS)_{jk}}} \quad (3.3)$$

$$F_{ik} \leftarrow F_{ik} \sqrt{\frac{(XGS^T)_{ik}}{(FF^T XGS^T)_{ik}}} \quad (3.4)$$

$$S_{ik} \leftarrow S_{ik} \sqrt{\frac{(F^T XG)_{ik}}{(F^T FSG^T G)_{ik}}} \quad (3.5)$$

通过式 (3.3)、(3.4)、(3.5) 对式 (3.2) 的反复迭代，可以使得式 (3.2) 达到最小值。

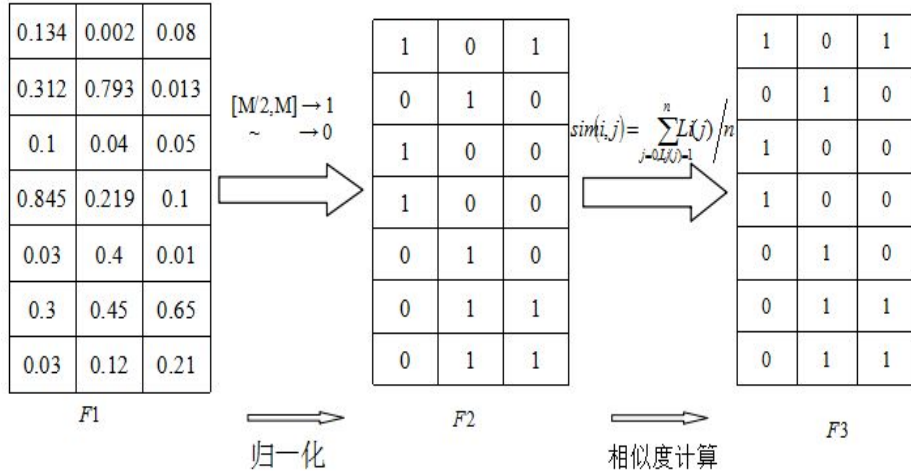


图 3.2 用户分类详情图

式 (3.2) 得到最优结果后, 就可以得到最终迭代好的 F , S , G 三个矩阵, 其中 F , G 分别表示用户和物品的特征信息。 F 和 G 为一些位于 0-1 之间的小数, 值越大, 意味着该因素对用户或者物品的影响因素越大, 即权重越大。在用户 (物品) 聚类之前, 需要对 F 和 G 进行预处理。详情如图 3.2 所示, 对于 F 中的每行 (G 中的每列), 值最大的小数记为 M , 在和 M 相同的小数数位下, 将 $[M/2, M]$ 之间的小数全部更新为 1, 其余更新为 0。这样, 就得到了能够表示用户或物品特征的行向量。预处理完成后, 根据式 (3.6) 来计算 F (G) 矩阵中行向量 (列向量) 的相似度。并且本文设置了一个相似度阈值 p , 相似度大于等于 p , 则认为属于同一类型用户 (物品)。通过每两个用户 (物品) 相似度的比较, 能够得到不同类型的用户 (物品)。其中, 可能有的用户既属于用户类型 X , 又属于用户类型 Y , 根据该用户与其他用户相似度的最高值进行归类。

$$sim(i, j) = \sum_{j=0, Lj(j)=1}^n Li(j) / n \quad (3.6)$$

式 (3.6) 为计算矩阵 i, j 行 (列) 向量的相似度, 其中 $Li(j)$ 表示第 j 行中为 1 的位置在第 i 行中的值。

表 3.1 用户分类算法

<p>Input: $M \times N$ 的辅助评分矩阵 X, 矩阵 S 的行数 K 和列数 L, 收敛值 \min, 判断行列向量相似度阈值 θ</p> <p>Output: 关于辅助评分矩阵 X 分类好的用户集合</p>
<pre> 1. 在[0, 1]之间随机初始化公式 3-1 中的 $F_{m \times k}^0$, $S_{k \times l}^0$, $G_{n \times l}^0$ 矩阵 2. for i = 1 ,..., I do 3. 利用公式 3-2, 3-3, 3-4 来迭代更新 $F_{m \times k}^{i-1}$, $S_{k \times l}^{i-1}$, $G_{n \times l}^{i-1}$, 得到最新的值 $F_{m \times k}^i$, $S_{k \times l}^i$, $G_{n \times l}^i$ 4. if 公式 3-1 的值 $\leq \min$ do 5. end for 6. end if 7. for i = 1,...,n do 8. $m = \max(F_i)$ 9. for j = 1,...,n do 10. if $(m, F_{ij}) \in$ 相同小数数量级 and $m/2 < F_{ij} < m$ do 11. $F_{ij} \leftarrow 1$ else $F_{ij} \leftarrow 0$ 12. end if 13. for i = 1,...,n do 14. for j = 1,...,n and j != i do 15. if 利用公式 3-5 得到 第 i 行和第 j 行的相似度 $\text{sim}(i,j) \geq \theta$ do 16. $[\text{userI}, \text{userJ}] \in$ 同一种用户类型 17. end if 18. end for 19. end for </pre>

3.3.2 基于 BP 神经网络的特征学习

在推荐系统中, 通常都会有对用户或者物品进行描述的标签, 而这些标签表示的就是用户或者物品的特征。通常情况下, 不同类型的用户(物品)的特征会不相同。比如女孩一般喜欢关注衣服、化妆品之类的物品; 而男孩一般关注电子、体育类的物品。因此, 可以根据特征的不同来判断该用户(物品)是属于哪种类型。

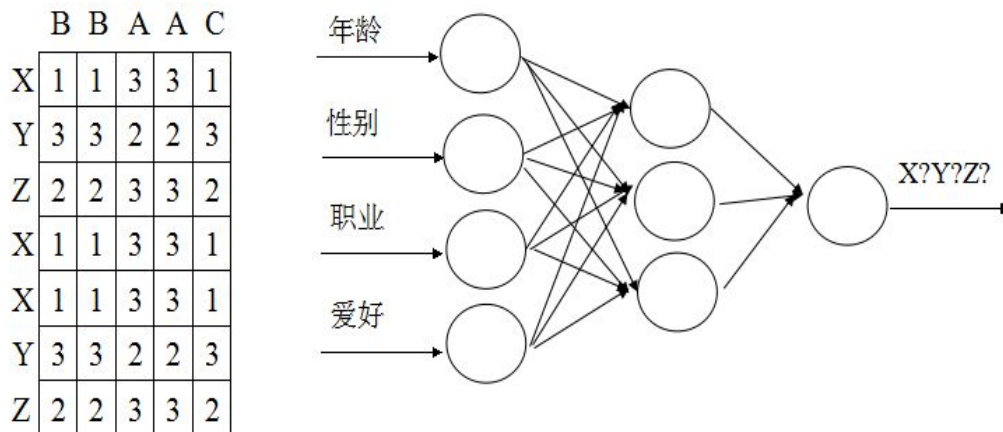


图 3.3 根据用户特征判断用户类型示意图

通过 3.2.1 中对用户的聚类，就能够得到属于不同类型的用户（物品），并且对这些相同类型的用户（物品）标签进行学习，就能够得到一个根据用户（物品）特征辨别它的类型的模型。本文采用 BP 神经网络进行特征学习，因为 BP 神经网络能够学习和存储大量的输入-输出模式映射关系（在本文中，这种映射关系表示为特征-类型），通过反向传播来不断调整网络的权值和阈值，使网络达到收敛。根据辅助域中用户的特征数以及收敛情况，本文建立一个输入层节点数为 n （ n 的大小为用户或物品的特征数），隐含层节点数为 $n1$ （ $n1$ 的确定参考式（3.6）），输出节点数为 m （ m 的大小为 1，即用户或物品的类型）的三层 BP 神经网络。如图 3.3 所示，比如年龄、性别、职业、爱好为用户的四个特征，X、Y、Z 为用户的类型。对用户（物品）特征和类型进行量化处理后，将用户的这四个特征作为神经网络的输入，用户类型作为输出，通过 BP 神经网络的若干次迭代便可以得到一条根据用户（物品）特征判断其类型的 BP 神经网络。

表 3.2 BP 神经网络类型预测值转换

$r \leq 0.05$	$0.05 < r \leq 0.15$	$0.15 < r \leq 0.25$	$0.25 < r \leq 0.35$	$0.35 < r \leq 0.45$	$0.45 < r \leq 0.55$
0	0.1	0.2	0.3	0.4	0.5
$0.55 < r \leq 0.65$	$0.65 < r \leq 0.75$	$0.75 < r \leq 0.85$	$0.85 < r \leq 0.95$	$r > 0.95$	
0.6	0.7	0.8	0.9	1	

在实验中，将辅助域的用户特征输入到训练好的神经网络时会得到一系列的位于 $[0, 1]$ 之间的输出值，需要对这些输出值进行一些处理以达到用户分类的较好效果，将该输出值用表 3.2 中的数据进行替换，然后根据值的不同将目标数据域的用户进行分类。


$$n_1 = \sqrt{n+m} + a \quad (3.7)$$

其中, n 为输入节点数, m 为输出节点数, 常数 $a=1\sim 10$ 。

3.3.3 对目标数据域的用户评分预测

目前, 大多数跨领域推荐模型主要是通过迁移其他相关领域的评分信息来解决目标域的数据稀疏性问题, 但由于评分刻度以及两个领域的不相关性等问题, 迁移的效果往往不是很理想。但如果不直接迁移评分信息, 而是将用户(物品)的特征信息作为两个领域迁移的连接点, 并且通过评分信息去填充空缺评分值。这样就能很好的解决上面的问题。虽然目标域的评分信息较为稀疏, 但用户(物品)的标签信息确非常丰富, 本文结合 3.2.2 中通过辅助域的数据训练好的神经网络以及目标域的数据中的用户(物品)标签信息来判别目标域中每个用户(物品)所属的分类。

	user /item	i1	i2	i3	i4	i5	i6	i7
X	u1	?	?	?	2	?	4	5
Y	u2	3	?	?	1	3	?	4
X	u3	2	3	?	?	5	?	?
Y	u4	2	?	4	?	5	5	5
Y	u5	?	2	3	?	?	?	4
X	u6	3	2	2	?	?	4	?



	user /item	i1	i2	i3	i4	i5	i6	i7
X	u1	2.5	2.5	2	2	4	4	5
Y	u2	3	2	3.5	1	3	5	4
X	u3	2	3	2	2	4	4	5
Y	u4	2	2	4	1	5	5	5
Y	u5	2.5	2	3	1	4	5	4
X	u6	3	2	2	2	4	4	5

图 3.4 填充目标数据域的空缺值

相同类型的用户对相同的物品的评分大致相同。根据这点, 只需要用相同类型的用户对同一物品的评分的平均值来填补同类型中没有对该物品评分的空缺值。这样, 就能够解决目标域中数据稀疏的问题。如图 3.4 所示, 目标评分矩阵中的‘?’代表空缺的评分值, 通过 3.2.1 中的用户聚类, 我们得到 $(u1, u3, u6) \in X$, $(u2, u4, u5) \in Y$, $u1$ 对 $i1$ 的评分为空, 而 $u3, u6$ 对 $i1$ 的评分分别为 $r(u3, i1)=2$, $r(u6, i1)=3$ 。由于 $u1, u3, u6$ 都属于同一类型 X , 所以只需要用 $u3$ 和 $u6$ 对 $i1$ 的平均评分来填充 $u1$ 对 $i1$ 的评分即可。因此, 目标域中的空缺评分值我们通过式 (3.8) 进行填充。

$$r(u_i, i_j) = \frac{\sum_{(u_x, i_j) \in G} r(u_x, i_j)}{n} \quad (3.8)$$

其中表示用户 i 对物品 j 的预测评分值, 表示和属于相同类型的并且已经有过评分的用户, n 表示的个数。

表 3.3 目标评分矩阵空缺值预测算法

<p>Input: $M \times N$ 的目标评分矩阵 Y, BP 神经网络迭代次数 $epochs$, 训练误差 $error$, 隐含层神经元个数 h。</p> <p>Output: 空缺处的预测值</p>
<pre>1. for t = 1,...,n (每行相同类型的用户) do 2. 用户特征 input (BP 神经网络输入) $\leftarrow [0, 1]$ 用户类型 output (BP 神经网络输出) $\leftarrow [0, 1]$ 3. end for 4. net.train(input,output,epochs,goal) // Train BP Neural Network 5. for g = 1,...,n do 6. tar_output = net.sim(tar_input) tar_outputs.append(tar_output) 7. end for 8. for line in tar_outputs do 9. 根据公式 3-3 对目标评分矩阵进行用户类别归类 10. end for 11. 使用公式 3-7 对目标评分矩阵空缺值进行预测并填充</pre>

3.4 实验评估

在这一小节中, 将通过真实的数据集来检验是否能够通过共同的标签信息, 将评分信息较为稠密的辅助域数据迁移到评分信息较为稀疏的目标域数据中去。在本次实验中, 主要从两点去评估算法的性能:

- 将本章提出的模型与当前的单领域推荐模型以及跨领域推荐模型进行比较, 验证该模型是否能解决协同过滤推荐系统中的数据稀疏问题, 能否更好提高推荐精度。
- 验证辅助矩阵评分稀疏度对算法的影响。

3.4.1 实验设计

在实验之前, 需要对实验中使用的数据集进行划分。并且实验还要考虑辅助评分矩阵稀疏性(不同的稀疏度会对辅助域的分类造成影响)对算法的影响。实验数据一共包含两种类别: 辅助评分矩阵和目标评分矩阵。将辅助评分矩阵全部划分为训练集, 将目标评分矩阵划分为 80%+20%两份, 其中 80%的数据用

户训练，20%的数据用于测试。根据稀疏度的不同将辅助评分矩阵进行了划分，用于分析稀疏度对算法的影响。每个数据域分成了 8 种数据，并且 8 组数据中的用户数和物品数一样，只是评分稀疏度不同，具体信息见表 3.4、表 3.5。

表 3.4 MovieLens 稀疏数据集

编号	用户数	电影数	评分比例	描述
1	250	250	10.5%	辅助矩阵
2	250	250	18.7%	辅助矩阵
3	250	250	25.2%	辅助矩阵
4	250	250	30.1%	辅助矩阵
5	250	250	38.4%	辅助矩阵
6	250	250	46.5%	辅助矩阵
7	250	250	50.3%	辅助矩阵
8	250	250	58.9%	辅助矩阵
9	300	300	8.9%	目标矩阵

表 3.5 Book-Crossing 稀疏数据集

编号	用户数	书籍数	评分比例	描述
1	250	250	15.8%	辅助矩阵
2	250	250	21.9%	辅助矩阵
3	250	250	25.6%	辅助矩阵
4	250	250	31.8%	辅助矩阵
5	250	250	39.7%	辅助矩阵
6	250	250	46.3%	辅助矩阵
7	250	250	50.1%	辅助矩阵
8	250	250	55.4%	辅助矩阵
9	300	300	15.7%	目标矩阵

在实验中，对于用户分类算法中的 k 和 l 的选择，选取的 k 和 l 分别为 15、15。对于判定用户相似度的 P 值，也选取了 0.65，经过 BP 神经网络的训练结果反馈，当 $\theta=0.8$ 时，用户归类最多。

3.4.2 实验数据及预处理

由于本文提出的方法是建立在两个领域拥有相同的标签的假设下，因此，把一个领域的数据集分成两个部分，一个部分的评分信息较为稠密，即辅助域数据，另一个部分的评分信息则较为稀疏，即目标数据域。对于该实验，选取

推荐系统中用于研究的公开的数据集来进行我们的实验评估，数据集如下：

- **MovieLens 数据集：**一个关于电影的评分数据集，它包含了超过 100000 个电影评分，总共有 943 个用户对 1682 个电影的评分，评分刻度为 1-5。并且这个数据集有对用户和物品进行描述的详细的标签信息。为了满足实验条件，对这个数据集进行辅助评分矩阵和目标评分矩阵的抽取。对于辅助评分矩阵，随机选取 250 个用户对 250 部电影的评分，并且将评分率人为控制在 10%-70%之间。对于目标评分矩阵，随机选取 300 个用户对 300 部电影的评分，通过抽样结果显示，抽取的目标评分矩阵评分率为 8.9%。
- **Book-Crossing 数据集：**一个关于图书的评分数据集，它包含了 1100000 个评分信息，总共有 278858 个用户对 271379 本书籍的评分，评分刻度为 0-9。同样的，这个数据集中也有对用户和物品进行描述的详细的标签信息。对于辅助评分矩阵，也随机选取 250 个用户对 250 本书籍的评分，并且将评分率人为控制在 10%-70%之间。对于目标评分矩阵，同样的随机选取 300 个用户对 300 本书籍的评分，通过抽样结果显示，抽取的目标评分矩阵评分率为 15.7%。

3.4.3 对比模型

由于本文提出的模型是通过迁移评分信息较为稠密的数据域的知识，来预测并填充目标数据域（也就是评分信息较为稀疏的评分矩阵）的空缺值。因此，本文主要和当今一些单领域的推荐模型以及跨领域的推荐模型进行比较。这些模型如下：

- **基于 NMF（non-negative matrix factorization）的模型：**这是一个单领域的推荐模型，它通过非负矩阵分解得到评分矩阵的隐含特征，并且根据这些特征信息去预测该评分矩阵的空缺值。
- **基于 CBS（Scalable cluster-based smoothing）的模型：**这是一个单领域的推荐模型，它通过计算同一集群中所有用户的可观察的评分值的平均值去预测并填充该评分矩阵的空缺值。
- **基于 CBT（Code Book Transfer）的模型：**这是一个跨领域推荐模型，它首先通过对辅助矩阵（评分信息较为稠密的评分矩阵）进行矩阵分解，从而得到 codebook（也就是所谓的知识），然后通过迁移该知识来预测并填充目标评分矩阵的空缺值。
- **FTLM（Feature Tags Learning Model）：**这是本章提出的跨领域推荐模型。

3.4.4 实验评估方法

本实验主要是预测目标评分矩阵的空缺值。因此，将采用 MAE (Mean Absolute Error) 来作为本实验的评价指标。计算公式如下：

$$MAE = \sum |r_i - r_i^*| / |O| \quad (3.9)$$

其中 O 表示测试集中需要预测的用户-物品评分信息，用 $|O|$ 表示 O 的大小。 r_i 表示评分的真实值， r_i^* 表示评分的预测值。MAE 值越小，代表预测结果越好。

3.4.5 实验结果及分析

根据表 3.4 和表 3.5 中的统计信息，分别在 MovieLens 数据集和 Book-Crossing 数据集上进行实验，通过实验得到的对比结果如下图 3.4, 3.5 所示。

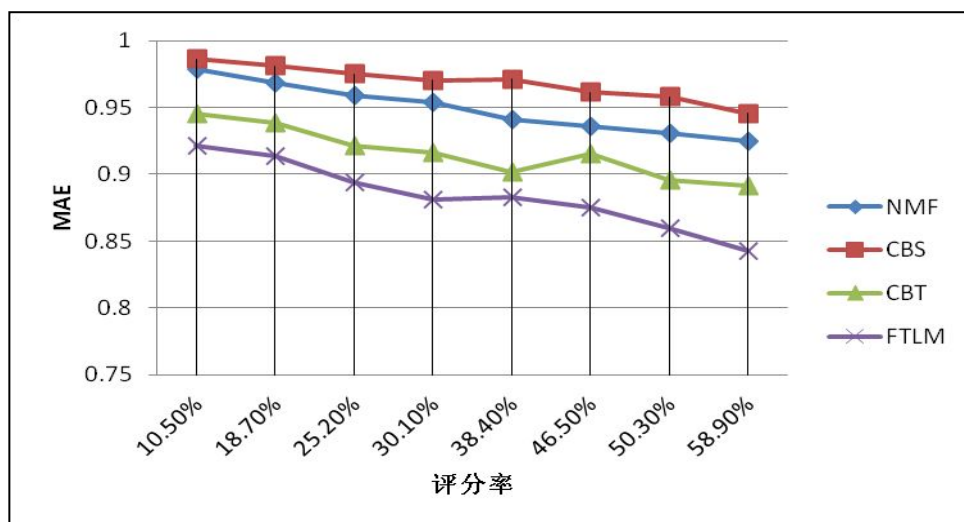


图 3.5 MovieLens 数据集中不同评分稀疏度的各模型的 MAE 值比较

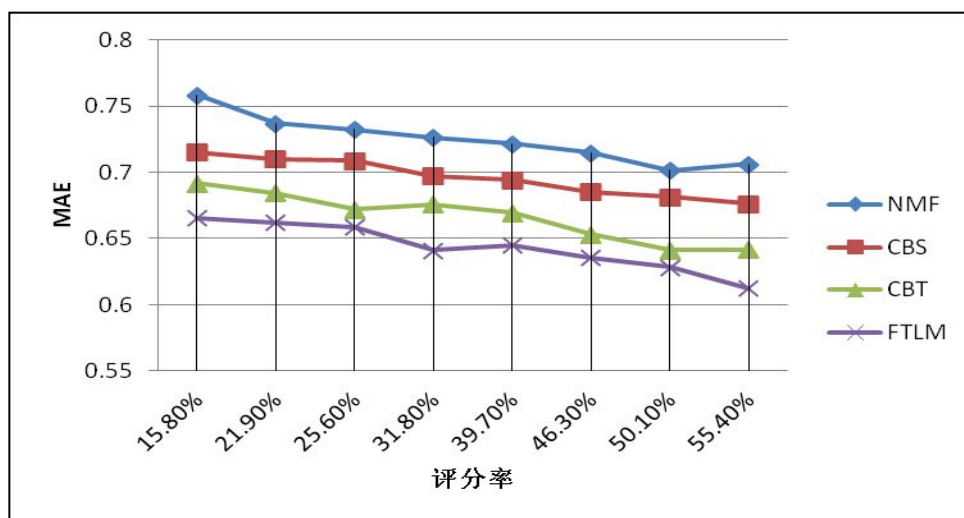


图 3.6 Book-Crossing 数据集中不同评分稀疏度的各模型的 MAE 值比较

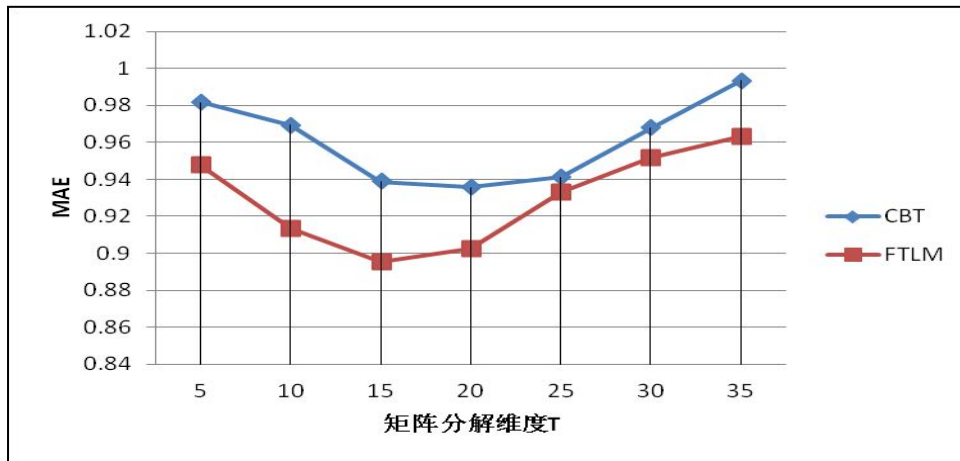


图 3.7 跨领域推荐模型中矩阵分解维度 T 对 MAE 值的影响

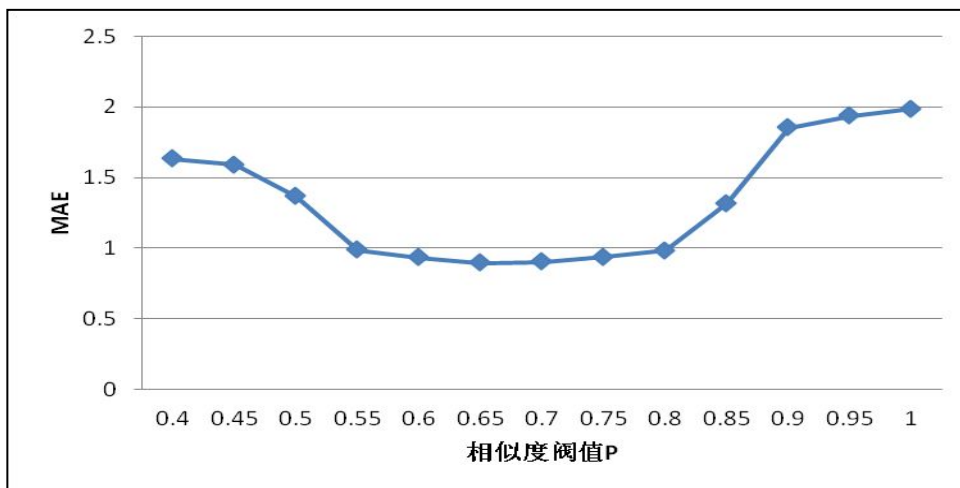


图 3.8 相似度阈值 P 对 MAE 值的影响

根据图 3.5, 3.6, 3.7, 3.8 的分析, 能够得到以下结论:

1) FTLM 模型和其他推荐模型对比:

通过图 3.5, 3.6 中的 MAE 和 RMSE 值的对比结果可以看到, 在相同的评分率的情况下, FTLM 模型比其他推荐模型 (图中 NMF、CBS、CBT) 有更低的 MAE 值, 这意味着 FTLM 模型能更好的解决目标评分矩阵的数据稀疏度这个问题。通过观察, 也能看到, 单领域推荐模型 (图中 NMF 和 CBS) 的 MAE 值普遍比跨领域推荐模型 (图中 CBT 和 FTLM) 的 MAE 值要高。可见, 通过不同领域之间的知识迁移还是能够较好的解决数据稀疏的问题, 提升推荐系统的性能。通过对各个模型的 MAE 值进行比较, 可以看到 CBS 模型比 NMF 模型有很低的 MAE 值, 这意味着基于聚类方法的模型能更好的挖掘出数据集中的隐藏信息, 能更加精确的预测评分空缺值。CBT 模型比 CBS 模型更优, 这又更好的证明了不同领域之间共享了一部分相同的评分信息, 通过共享的评分能够更好的填充目标领域空缺的评分值。在跨领域推荐模型 CBT 和 FTLM 模型的

对比中可以得到，虽然跨领域推荐都能够更好的解决数据稀疏这个问题，但 CBT 是通过迁移不同领域之间的评分信息来填充目标评分矩阵的空缺值，但两个领域并不一定处处都是正相关的，有可能某一部分的评分信息并不能迁移到另一领域，强行迁移会造成数据的负迁移。而 FTLN 模型是通过学习辅助域中用户和物品的特征（通过标签信息进行学习），找到不同类型的用户和物品特征参数，将辅助域中学到的信息运用到目标域中，利用目标域中用户和物品特征结合通过辅助域学得到的模型来预测并填充空缺值，这就很好的解决了两个领域之间的差异性造成的负迁移效果。

2) 辅助评分矩阵的不同稀疏度对算法的影响评估：

从图 3.5, 3.6 中可以明显的看到，随着评分率的下降（数据稀疏度的上升），各个模型的 MAE 和 RMSE 值均呈现出上升的趋势，意味着数据稀疏度越大，预测的结果越不精确。这也很容易理解，辅助评分矩阵如果越稀疏，那么基于矩阵分解技术的这些模型再求最优解时就很容易造成过度拟合的情况，这样就造成了我们的结果不准确。

3) 分解维度 T 和相似度阈值 P 对 MAE 值的影响：

从图 3.7 可以看到 MAE 值变化的整体趋势，MAE 值刚开始随维度 T 的增大而下降，当达到某一特定值时，这时 MAE 值最小，之后便是随维度 T 的增大而增大。从 FTLN 的曲线图上可以看到，当 $T=15$ 时，MAE 值达到最小。在 3.3.1 中，设有一个相似度阈值用于判断用户的类型，不同的阈值 P 会得到不同类型的用户分组。从图 3.8 中可以看到，MAE 随着相似度的阈值先增大而增大，当相似度阈值 P 达到 0.65 时，MAE 值最小。随后又随着 P 的增大而减小。

通过以上实验及分析表明，可以通过迁移学习来解决协同过滤推荐系统中的数据稀疏问题，但在选择辅助域的数据时，还要考虑辅助数据域的稀疏度等问题，应当尽量选取哪些数据稀疏度不高的辅助数据域。这样，才能更好的进行迁移，预测并填充目标评分矩阵。

3.5 本章小结

本章主要介绍了一种基于标签的跨领域推荐方法。该方法利用标签来连接两个领域，避免了领域间的评分刻度差异所带来的问题，解决了领域间的差异性导致的数据负迁移。该方法首先通过基于非负矩阵正交分解聚类算法对辅助评分矩阵进行分类，得到不同类型的用户或物品。然后，通过 BP 神经网络对不同类型的用户（物品）的特征（用标签进行描述）进行学习，训练得到根据用户的特征判断该用户类型的 BP 神经网络。最后，本文通过目标域中的用户

或物品的特征对目标域中的用户或物品进行分类，并且用同类型的平均评分去填充那些空缺的评分。在实验中，通过与单领域和跨领域的推荐模型进行比较，验证了本文提出的 FTLM 模型能更好的去解决目标评分矩阵中的数据稀疏问题。同时，通过不同数据稀疏度下 MAE 的比较，验证了数据稀疏度对算法的影响，即数据稀疏度越高，算法的性能越低。

第4章 基于潜在特征聚类的跨领域推荐方法

4.1 概述

在第三章中，通过额外的标签信息完成了不同领域之间的知识的迁移，但目前不同领域之间具有相同类型的标签这个假设还是比较难以达到的。因此，在本章中，主要介绍通过潜在的特征来完成跨领域推荐任务。大多数推荐系统都是基于协同过滤技术的，它是依靠用户对一系列物品的历史评分记录去为用户推荐物品。然而在现实中，用户仅仅会对一部分物品进行评分，使得评分信息非常稀疏，这就使得推荐模型在计算时会遇到过度拟合问题，从而造成推荐结果不精确的问题。事实上，不同领域之间会存在共有的评分信息，用户的偏好可以从一个领域迁移到另一领域。最近，有一些跨领域的推荐模型被提出^[14,19]，这些模型通过从稀疏度较低的领域抽取“知识”，然后将抽取得到的“知识”迁移到目标领域中去。这些推荐模型有两个主要的限制条件：

1) 大多数推荐模型都是从另一个领域抽取一个共有的潜在评分模式作为知识进行迁移。但在实际情况下，不同领域之间是否相关并不知道，所以这个抽取的潜在的评分模式并不一定能代表两个领域的评分特性。

2) 即使两个相关的领域确实共享了一个共有的潜在评分模式，但这些领域之间也是有一定的差异性，这些差异性的影响可能大于共有的潜在评分模式，这会造成更差的推荐结果。也就是说，现有的推荐模型都没考虑领域间评分模式的那些特殊知识。

因此，在本章中，提出一个改进的基于潜在特征聚类的跨领域推荐方法。该方法主要研究了不同领域之间的相关性，通过领域之间的相关性来确定领域之间的共享维度和私有维度，共享维度可以确定领域之间所共享的知识，这一部分知识用来进行迁移。而私有维度则代表各自领域的私有特性，这一部分知识是用来更好的还原每一个领域的特性。也就是说改进后的基于潜在特征聚类的跨领域推荐模型不仅会对这部分共享的知识进行学习。同时，还会对领域之间那些特殊的知识进行学习，这样通过每个领域的特殊知识就能消除领域间的不相关性所带来的数据负迁移问题。

4.2 问题定义

假设对多领域的个性化物品进行推荐, τ 表示领域的索引, $\tau \in [1, t]$ 。在第 τ 个领域的评分矩阵 D_τ 中, 有用户集合 $X_\tau \in \{X_1^\tau, \dots, X_{M_\tau}^\tau\}$ 对物品集合 $Y_\tau \in \{Y_1^\tau, \dots, Y_{N_\tau}^\tau\}$ 进行评分。其中, M_τ 代表行数 (用户), N_τ 代表列数 (物品)。不同领域间的用户集合和物品集合可能重叠也可能独立。并且每个评分矩阵都有一些评分值和空缺值。因此, 还设有一个二进制的权重矩阵 W_τ , 该权重矩阵的大小和评分矩阵 D_τ 一致, 如果 $[D_\tau]_{ij}$ 有评分值, 则 $[W_\tau]_{ij} = 1$; 否则 $[D_\tau]_{ij} = 0$ 。和第三章一样, 将评分信息较为稀疏的领域, 也就是需要预测空缺评分值的领域称作目标领域, 对应的评分矩阵称为目标矩阵; 将评分信息较为稠密的领域称为辅助域, 对应的评分矩阵称为辅助矩阵。

4.3 基于潜在特征聚类的跨领域推荐方法

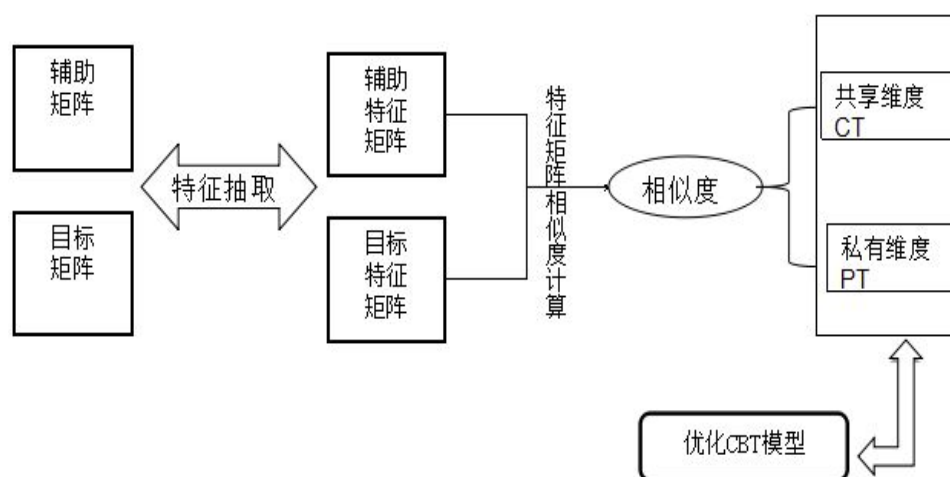


图 4.1 基于潜在特征聚类的跨领域推荐框架

图 4.1 是基于潜在特征聚类的跨领域推荐框架图, 主要包括以下几个流程: 1) 对辅助矩阵和目标矩阵进行特征抽取, 以得到它们各自的特征矩阵; 2) 对辅助特征矩阵和目标特征矩阵进行特征矩阵的相似度计算以得到它们之间的相似度; 3) 根据得到的相似度确定两个领域的共享知识的维度和特殊知识的维度; 4) 根据共享维度和私有维度对 CBT 模型进行优化。

4.3.1 CBT 模型

B.Li 在^[14]中提出了一个跨领域推荐模型, 他们首先对用户和物品进行聚类, 抽取出领域间所共有的评分模式 (共享的知识)。在领域 τ 中的目标矩阵的初始聚类可以用非负正交矩阵分解模型进行优化, 优化公式如下:

$$\min_{U_\tau, S_\tau, V_\tau \geq 0} \lambda_\tau = \left\| [D_\tau - U_\tau S_\tau^* V_\tau^T] \circ W_\tau \right\|^2 \quad (4.1)$$

其中, $U_\tau \in R^{M_\tau \times K_\tau}$ 表示第 τ 领域的 K_τ 用户聚类; $V_\tau \in R^{N_\tau \times L_\tau}$ 表示表示第 τ 领域的 L_τ 物品聚类; $S_\tau^* \in R^{K_\tau \times L_\tau}$ 代表 k 个用户聚类对 l 个物品聚类的评分模式。 W_τ 表示二进制掩码矩阵; \circ 表示两矩阵对应相乘。

上述模型从辅助领域中抽取知识, 然后全部迁移到目标领域中。如上所说, 它并没有考虑到两个领域之间的相关性。但是, 不同领域之间还存在一些特殊的知识, 本章提出的方法不仅要学习这些共有的知识, 还要学习那些特殊的知识。所以本文将不同领域之间的潜在评分模式分成公有部分和特殊部分, 也就是将 S_τ^* 分成两份, 即 $S_\tau^* = [S_0, S_\tau]$, 其中 $S_0 \in R^{K_\tau \times T}$, $S_\tau \in R^{K_\tau \times (L_\tau - T)}$, T 是两个领域间共有评分模式的维度, $(L_\tau - T)$ 是第 τ 领域间特殊的评分模式。 S_0 表示两个领域间共有的知识, 通过迁移这部分知识我们能够减轻目标域的数据稀疏性问题。 S_τ 表示两个领域之间的差异性, 它能很好的反应领域间的相关度, 并且能更好的提高预测精度。

如图 4.2 所示, U_1, U_2 表示用户聚类的潜在特征矩阵, V_1, V_2 表示物品聚类的潜在特征矩阵, S_0 表示两个领域共有的知识, S_1, S_2 分别表示辅助和目标领域的特殊知识。CLFM 模型学习用户聚类潜在特征 $U_\tau \in R^{M_\tau \times K_\tau}$ 和物品聚类潜在特征 $V_\tau = [V_{\tau 0}^T, V_{\tau 1}^T] \in R^{N_\tau \times L_\tau}$ 。

因此, 本章提出的模型可以用下面公式作为目标函数:

$$\min_{U_\tau, S_\tau, V_\tau \geq 0} \lambda_\tau = \sum_\tau \left\| [D_\tau - U_\tau [S_0, S_\tau] V_\tau^T] \circ W_\tau \right\|^2 \quad (4.2)$$

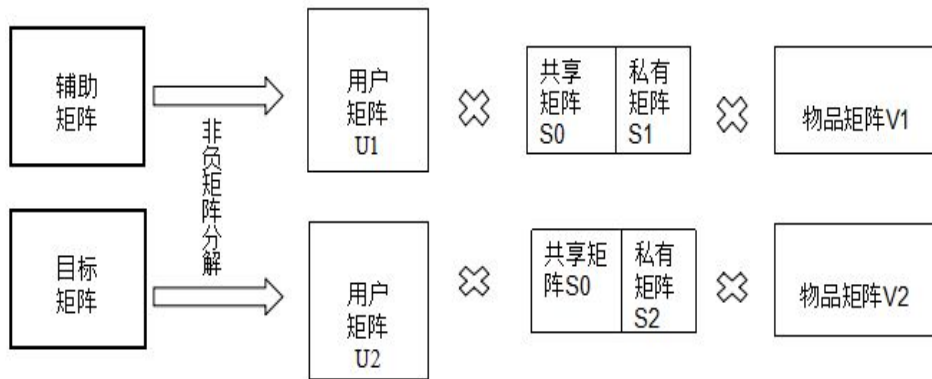


图 4.2 CLFM 模型图示说明。

4.3.2 CBT 模型优化

本文提出的模型可以用交替最小化算法进行优化, 式(4.2)中的目标函数可以改写成如下形式:

$$\min_{U, S_0, S_1, S_2, V \geq 0} \lambda = \left\| [D_1 - U_1[S_0, S_1]V_1^T] \circ W_1 \right\|^2 + \left\| [D_2 - U_2[S_0, S_2]V_2^T] \circ W_2 \right\|^2 \quad (4.3)$$

其中 $U_1 \in R^{M_1 \times K}$, $U_2 \in R^{M_2 \times K}$, $V_1 = [V_{10}^T, V_{11}^T] \in R^{N_1 \times L_1}$, $V_2 = [V_{20}^T, V_{21}^T] \in R^{N_2 \times L_2}$, $S_0 \in R^{K \times T}$, $S_1 \in R^{K \times (L_1 - T)}$, $S_2 \in R^{K \times (L_2 - T)}$ 。

采用交替乘法更新算法^[61]来优化本文提出的模型。对于 U , S_0 , S_1 , S_2 , V 式 (4.3) 中目标函数 λ 是非凸的, 因此交替乘法更新算法通过固定一个参数, 然后不断修改其他参数来使目标函数达到最优值, 重复迭代直到达到收敛。推导 S_1 : 将式 (4.3) 改写成如下形式, 根据前面的描述, D_1 表示源领域, D_2 表示目标领域:

$$\min \lambda(S_1) = \left\| [D_1 - U_1 S_0 V_{10} - U_1 S_1 V_{11}] \circ W_1 \right\|^2 + \left\| [D_2 - U_2 S_0 V_{20} - U_2 S_2 V_{21}] \circ W_2 \right\|^2 \quad (4.4)$$

4.3.3 领域相关性分析

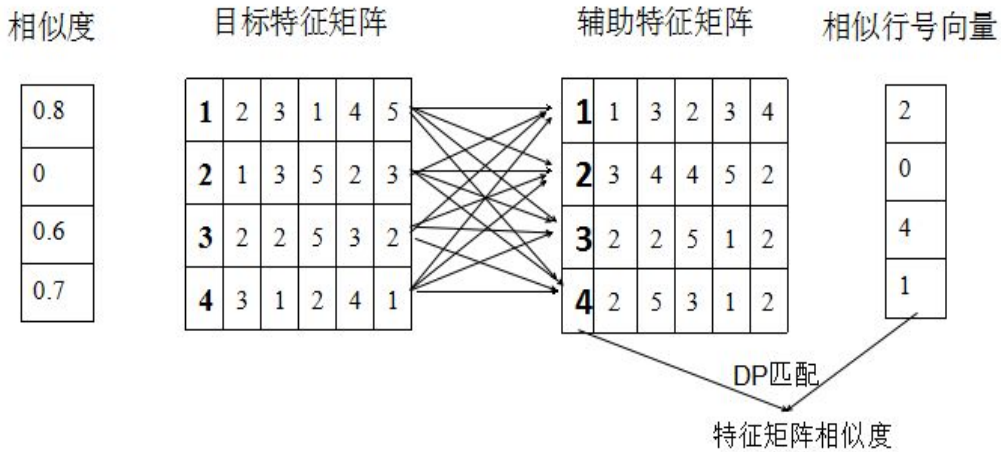


图 4.3 基于 DP 匹配的特征矩阵相似度判断

将不同领域之间的潜在评分模式分成共享部分和私有部分, 也就是将 CBT 中的知识 S_r^* 分成两份, 即 $S_r^* = [S_0, S_r]$ 。但我们并不知道共享部分和私有部分的维度, 在这一节中主要分析领域的相关性, 计算不同领域间的相似度, 通过领域间的相似度来确定共享维度和私有维度。首先, 抽取辅助矩阵和目标矩阵的特征矩阵, 然后通过 DP 匹配距离算法计算两个特征矩阵的距离, 进而得到特征矩阵之间的相似度, 该相似度即代表领域间的相似度。

1) 特征矩阵抽取

首先要对辅助域的数据矩阵以及目标域的数据矩阵进行特征抽取, 抽取的特征矩阵要能代表两个领域内各自数据的特点。因此, 本文采用 Li 在^[14]中提出的码书构建算法 (Codebook Construction Algorithm), 该算法通过对原始矩阵

的聚类压缩得到一个码书（Codebook），该码书是一个 $k \times l$ 的矩阵，代表 k 类用户对 l 类物品的评分规则，该矩阵中的每一个评分都能代表原始矩阵中的若干用户对若干物品的评分。所以，码书能够很好的代表原始矩阵的特征。

2) 基于 DP 匹配的特征矩阵相似度计算

通过码书构建算法分别得到了辅助域和目标域的 $k \times l$ 特征矩阵，特征矩阵是能够代表每个领域的特征，如果特征矩阵相同或者相似，就能判定两个领域是相关的；反之，则两个领域不相关。欧氏距离能够判断向量的空间位置距离，进而能判断向量之间的相似度，但特征矩阵经过个行列向量的转置位移等，特征矩阵中的每个点的信息也不是一一对应的关系，所以传统的判定向量距离的算法不适用特征矩阵相似度的判断。而 DP（Dynamic Programming）匹配距离算法也是一种判断向量距离的方法，和传统的欧几里得范数不同，DP 匹配算法有了一个滑动窗口的概念，允许特征向量中的一个元素可以在给定的滑动窗内选择对应特征向量中具有最小元素距离的元素作为其匹配对象，而不仅仅限定在对应位置元素。设整合窗的大小为 R ，第 i 个向量元素可以在对应特征向量 $[i-R, i+R]$ 范围的元素中选择最适当的向量元素，从而解决特征向量间的非精确匹配问题。如图 4.3 所示，目标特征矩阵的行向量 i 与辅助特征矩阵中的行向量 j 进行 DP 匹配，行向量中每个元素的距离公式如下：

$$d(I_{\alpha i}, Q_{\beta j}) = \frac{|I_{\alpha i} - Q_{\beta j}|}{I_{\alpha i} + Q_{\beta j}} \quad (4.5)$$

其中，分别代表目标特征矩阵中第 α 行第 i 个元素和辅助特征矩阵中第 β 行第 j 个元素。行向量 α 和行向量 β 的 DP 匹配距离就是求 α ， β 中每个元素的距离组成的有向图的最短路径，通过 Dijkstra 最短路径算法就可以求得，最短路径记为，当满足式（4.6）时，则将行号记录到相似行号向量 SL 中（当 $\alpha_1 \neq \alpha_2$ ， $SL[\alpha_1] \neq SL[\alpha_2]$ ），若不满足，则在相似行向量的相应位置插入 0。

$$SL[\alpha] = \begin{cases} k, d_{\min}(\alpha, k) < Y, Y \text{ 为阈值} \\ 0, \text{相反} \end{cases} \quad (4.6)$$

对目标特征矩阵的所有行向量进行一维 DP 匹配后，得到了目标特征矩阵的相似行向量，再相似行向量和以辅助特征矩阵行号为元素组成的向量进行 DP 匹配距离计算。就可以得到两个特征矩阵的匹配距离，并对该距离进行归一化处理，如式（4.7），忽略了向量中元素个数的影响。

$$d_{DP}(\alpha, \beta) = \frac{1}{2n-1} (d(1,1) + P_{\min}) \quad (4.7)$$

则两个特征矩阵的相似度为 $sim(\alpha, \beta) = 1 - d_{DP}(\alpha, \beta)$ 。即共享维度/私有维度 $= sim(\alpha, \beta) / (1 - sim(\alpha, \beta))$ 。

4.4 实验评估

在这一小节中，将通过真实的数据集来检验是否能够通过共同的标签信息，将评分信息较为稠密的辅助域数据迁移到评分信息较为稀疏的目标域数据中去。在本次实验中，主要从两点去评估算法的性能：

- 将本章的模型与当前的单领域推荐模型以及跨领域推荐模型进行比较，验证该模型是否能解决协同过滤推荐系统中的数据稀疏问题，能更好提高推荐精度。
- 分析本章模型是否比 CBT 模型更能反应知识迁移的真实情况。

4.4.1 实验数据

本次实验和章节 3 中的实验不同，不需要人为分割数据集成辅助域和目标域，而是通过现实存在的两个数据集进行评估。对于该实验，选取推荐系统中用于研究的公开的数据集来进行实验评估，数据集如下：

- **MovieLens 数据集：**一个关于电影的评分数据集，它包含了超过 100,000 个电影评分，总共有 943 个用户对 1,682 个电影的评分，评分刻度为 1-5。随机选取了 500 个用户对 1000 部电影作为本次实验的数据，每个用户至少有 16 个评分记录。
- **EachMovie 数据集：**一个关于电影的评分数据集，它包含了 2.8 百万个电影评分，总共有 72,916 个用户对 1,628 个电影的评分，评分刻度为 1-6。随机选取了 500 个用户对 1000 部电影作为本次实验的数据，每个用户至少有 20 个评分记录。
- **Book-Crossing 数据集：**一个关于图书的评分数据集，它包含了 1.1 百万个评分信息，总共有 278,858 个用户对 271,379 本书籍的评分，评分刻度为 0-9。同样的，随机选取了 500 个用户对 1000 部电影作为本次实验的数据，每个用户至少有 16 个评分记录。

4.4.2 对比模型

由于 CLFM 模型是通过迁移评分信息较为稠密的数据域的知识，来预测并填充目标数据域（也就是评分信息较为稀疏的评分矩阵）的空缺值。因此，本章主要和一些当今单领域的推荐模型以及跨领域的推荐模型进行比较。这些模型如下：

- **基于 NMF（non-negative matrix factorization）的模型：**这是一个单领域的推荐模型，它通过非负矩阵分解得到评分矩阵的隐含特征，并且根据这些特征信息去预测该评分矩阵的空缺值。
- **基于 FMM（Flexible Mixture Model）的模型：**这是一个单领域的推荐

模型，它使用概率混合模型学习每个领域的潜在聚类结构，然后分别提供单领域的性能。

- 基于 CBT (Code Book Transfer) 的模型：这是一个跨领域推荐模型，它首先通过对辅助矩阵(评分信息较为稠密的评分矩阵)进行矩阵分解，从而得到 codebook (也就是所谓的知识)，然后通过迁移该知识来预测并填充目标评分矩阵的空缺值。
- CLFM (Cluster-Level Factor Model)：这是本章提出的跨领域推荐模型。

4.4.3 实验设计

对于每个数据集，随机选取 300 个用户的评分作为训练集，剩下的那 200 个用户的评分则作为测试集。在本次实验中，对于每个测试集，设置不同大小的可观察的用户评分初始化，比如 5、10 或者 15 个评分信息（表 4.1 中显示的 ML-Given5 或 ML-Given10 或 ML-Given15），测试集中剩下的评分信息则用来评估。

选择 MovieLens vs EachMovie, EachMovie vs Book-Crossing 和 MovieLens vs Book-Crossing 作为 3 种相关领域（对于 CBT 模型来说，前面的是辅助域，后面的是目标域）。

4.4.4 实验评估方法

我们的实验主要是预测目标评分矩阵的空缺值，因此我们将采用 MAE (Mean Absolute Error) 作为我们实验的评价指标。计算公式如下：

$$MAE = \sum |r_i - r_i^*| / |O| \quad (4.8)$$

其中 O 表示测试集合中需要预测的用户-物品评分信息，我们用 $|O|$ 表示 O 的大小。 r_i 表示评分的真实值， r_i^* 表示评分的预测值。MAE 值越小，代表预测结果越好。

表 4.1 各模型在 MovieLens 和 EachMovie 对比中的 MAE 值

		NMF	FMM	CBT	CLFM
MAE	ML-Given5	0.9772	0.9456	0.9321	0.9241
	ML-Given10	0.9584	0.9314	0.9237	0.8934
	ML-Given15	0.9471	0.9305	0.9195	0.8827
	EM-Given5	0.9951	0.9672	0.9401	0.9213
	EM-Given10	0.9536	0.9410	0.9241	0.8936
	EM-Given15	0.9503	0.9362	0.9158	0.8897

表 4.2 各模型在 EachMovie 和 Book-Crossing 对比中的 MAE 值

		NMF	FMM	CBT	CLFM
MAE	EM-Given5	0.9836	0.9636	0.9645	0.9297
	EM-Given10	0.9470	0.9421	0.9344	0.9144
	EM-Given15	0.9386	0.9309	0.9321	0.9089
	BC-Given5	0.7485	0.7251	0.6996	0.6874
	BC-Given10	0.7238	0.6988	0.6879	0.6418
	BC-Given15	0.7215	0.6893	0.6826	0.6364

4.4.5 实验结果及分析

在这一小节中，将比较不同模型在不同参数下的 MAE 值。在每个评分数据集中，用户聚类 K 和物品聚类 L 在 $[10, 100]$ 之间选取。对 NMF 模型，矩阵分解的维度设置为 50。对 CBT 模型，将 K 设置为 20， L 设置为 35。

表 4.1 是不同的推荐模型在 MovieLens 和 EachMovie 数据集上的 MAE 比较结果。在本次实验中，有 5、10 和 15 个观察的评分值用于训练，剩下的评分值则用户评估。从实验结果可以看到最好的模型是本章提出的 CLFM 模型。FMM 模型的结果比 NMF 模型略好，这意味着基于聚类方法的模型能更好的挖掘出数据集中的隐藏信息，能更加精确的预测评分空缺值。并且还能清楚的看到，单领域推荐模型（如 NMF 和 FMM）的 MAE 值普遍比跨领域推荐模型（如 CBT 和 CLFM）的 MAE 值要高，可见，通过不同领域之间的知识迁移还是能够较好的解决数据稀疏的问题，提升推荐系统的性能。并且本章提出的 CLFM 模型比当今跨领域推荐模型的性能更好，也证明了该模型通过抽取共有的知识和特殊的领域知识比纯粹的迁移共有的知识符合实际情况，更加能够提高跨领域推荐的准确度。

同样的，表 4.2 是不同的推荐模型在 EachMovie 和 Book-Crossing 数据集上的 MAE 比较结果。从这个实验结果也能得出相似的结论，本章提出的 CLFM 模型比其他推荐模型更加优秀。

然而，从表 4.1 和表 4.2 中，还能够发现当 EachMovie 和不同领域的评分数据集进行实验时，实验结果也不相同，即时是相同的用户和相同的物品。这个结果意味着，不同领域之间所共享的知识是不相同的，而 CLFM 模型的显著优势是可以控制共享空间维度 T 。在该模型中，参数 T 被控制在 0 和 $\min(L1, L2)$ 之间，比如两个领域完全没共享知识和共享知识完全重叠。

图 4.4 显示了在 $K=20$ ， $L=35$ ，给出 10 个可观察的评分值的条件下，各个模型在 EachMovie 数据集下关于最短路径阈值 Y 的 MAE 值变化趋势图。从该

图中可以看到, 当 $Y < 0.7$ 的时候, 该模型的 MAE 值随 Y 增大而减小, 当 $Y = 0.7$ 的时候, 此刻的 MAE 值最低, 当 $Y > 0.7$ 时, MAE 值则随 Y 的增大而增大。

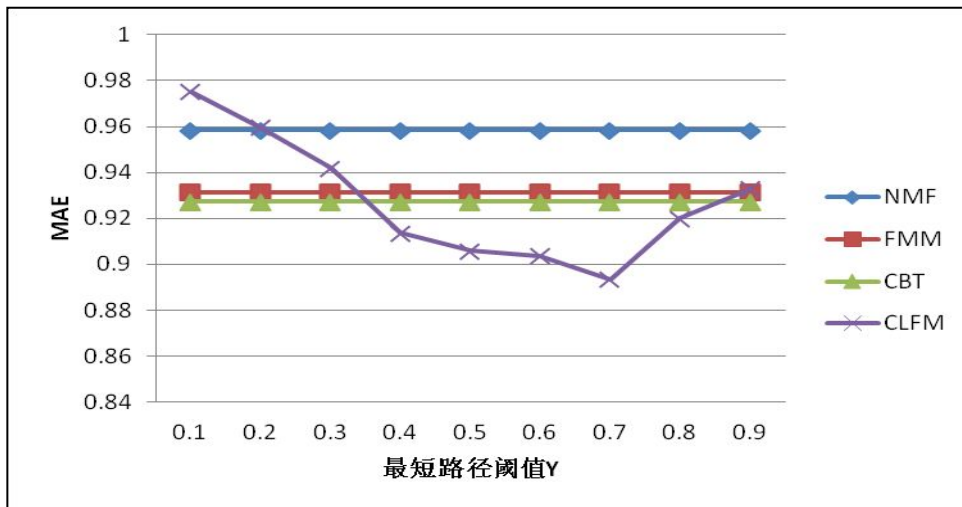


图 4.4 在 EachMovie 数据集上, 推荐模型关于最短路径阈值 Y 和 MAE 值的关系图

4.5 本章小结

本章主要介绍了基于潜在特征聚类的跨领域推荐方法。该方法不仅学习不同领域间的“共享知识”, 同时也学习各自领域的“特殊知识”, 通过这部分“特殊知识”去消除领域间的不相关性所带来的数据负迁移问题。并且, 本文还通过 DP 距离匹配算法计算两个领域的相似度, 通过该相似度来得到共享知识矩阵的维度以及特殊知识矩阵维度, 提高方法的效率。本文首先提出了该方法的核心思想。然后阐述了模型的具体优化过程。最后, 通过实验将本文提出的方法与单领域推荐模型、跨领域的推荐模型进行比较, 验证了 CLFM 模型能更好的去解决目标评分矩阵中的数据稀疏问题, 并且比当前的跨领域推荐模型有更好的性能。

第 5 章 总结与展望

5.1 工作总结

本文首先对传统推荐系统的产生进行了一个简单回顾，并阐述了单领域推荐系统中所存在问题，然后引出了迁移学习和跨领域推荐这两个概念，并对迁移学习和跨领域推荐的当今研究现状做了一番陈述。其中重点介绍了迁移学习和跨领域推荐中的一些模型。在针对单领域推荐系统存在的问题上，提出了两种基于迁移学习的跨领域推荐方法来解决这些问题。本文的贡献主要包括以下两个方面：

1) 介绍了一种基于标签的跨领域推荐方法，该方法用于解决协同过滤推荐系统中数据稀疏性问题。首先通过基于非负矩阵正交分解聚类算法对辅助评分矩阵进行分类，这样，能够得到不同类型的用户或物品。然后，建立了一个 BP 神经网络对不同类型的用户或物品的特征（用标签进行描述）进行学习，得到一条通过用户的特征判断该用户类型的 BP 神经网络。最后，通过目标域中的用户或物品的特征对目标域中的用户或物品进行分类，并且用同类型的平均评分去填充那些空缺的评分。在实验中，通过与单领域和跨领域的推荐模型进行比较，验证了 FTLM 模型能更好的去解决目标评分矩阵中的数据稀疏性问题。同时，通过不同数据稀疏度下 MAE 的比较，验证了数据稀疏度对算法的影响，即数据稀疏度越高，算法的性能越低。

2) 介绍了一个基于潜在特征聚类的跨领域推荐方法。该方法不仅学习不同领域间共享的评分模式（共享知识），同时也学习各自领域的特殊评分模式（特殊知识）。这样既能减轻目标域的数据稀疏性问题，同时也能消除不同领域间的差异性。首先提出了我们模型算法的核心思想。然后阐述了模型的具体优化。最后，在实验中，通过与单领域和跨领域的推荐模型进行比较，验证了 CLFM 模型能更好的去解决目标评分矩阵中的数据稀疏性问题，并且比当前的跨领域推荐模型有更好的性能。同时，在实验中还验证了共享空间维度 T 和 MAE 值的关系，了解到维度 T 对 MAE 值的影响，当 T 达到某一特定值时，MAE 值最小，模型效果最好。也通过实验证明了 CLFM 模型能较快达到收敛，保证了模型在跨领域推荐任务中的效率。

5.2 进一步工作及展望

虽然本文利用标签学习在跨领域推荐算法上进行了一定的研究，并且也取得了比较好的实验结果，得出了实验结论。但本文提出的两种基于迁移学习的跨领域推荐方法还有很多可以进行优化和改进的地方。

针对第三章提出的基于标签的跨领域推荐方法，有两点可以进行改进。首先，FTLM模型需要两个领域共享相同的标签。但在实际情况下，两个领域用来描述用户和物品的标签通常各不相同，可以去研究标签信息之间的相关性，用不同领域间的相似的标签来作为知识迁移的桥梁，让FTML模型的实际意义更大。其次，在对目标数据域进行用户评分预测过程中，只是取相同类型的用户对相同物品的评分的平均值来填充空缺的评分，可以再对相同类型的用户与未评分用户的关系再进行进一步分析，为未评分用户的评分值与其他用户的评分值确定一个权重系数，让预测评分值更加接近真实值。

对于第四章提出的基于潜在特征聚类的跨领域推荐方法，还可以进一步充实和优化。首先，在更大规模的评分数据集上来验证CLFM模型的计算能力。其次，在处理不同领域的评分刻度差异问题上，不应简单的对评分值进行统一处理，应该根据每个领域的实际情况进行处理。

致 谢

将近三年的研究生学习生活已接近尾声，回想这三年，一切历历在目。在这三年里，我收获了很多，懂得了研究生的意义所在。一路走来，虽然坎坷众多，但让我学会了如何面对困难，如何处理困难。即将离开待了快三年的实验室也让我不舍，在实验室里做工程、写论文；在实验室里被老师鼓励、批评，一切的一切都是我以后生活中的美好记忆，虽然有痛有累，但我依旧享受。我要感谢在这三年的时间里陪伴我的老师，同学。因为你们，让我研究生的学习生活充满了乐趣。

感谢我的导师万健教授。万老师留给我最深的印象是务实和风趣，最影响我的一句话是如果不比老师花更多的时间去学习，那在成就上就永远赶不上老师。由衷的祝福万老师工作顺利，在工作之余也要注意身体。

感谢殷昱煜老师。殷昱煜老师为人谦和，平易近人。在论文的选题，搜集资料和写作阶段，殷昱煜老师都倾注了极大的关怀和鼓励。在论文的写作过程中，每当我有所疑问，殷昱煜老师总会放下繁忙的工作，不厌其烦地指点我；在我初稿完成之后，殷昱煜老师又在百忙之中抽出空来对我的论文认真的批改，字字句句把关，提出许多中肯的指导意见，使我在研究和写作过程中不致迷失方向。他严谨的治学之风和对事业的孜孜追求将影响和激励我的一生，他对我的关心和教诲我更将永远铭记。

感谢张纪林老师。张老师平常风趣幽默，但在处事上还是非常认真和严格的。这种严格让我们实验室的每一位都吃尽苦头，虽然稍许抱怨，但我们能够体会张老师的良苦用心，张老师这种严谨处事的态度也会影响我的一生。

感谢云技术研究中心的蒋从锋老师、任祖杰老师、张伟老师、司华友老师、游新冬老师、周仁杰老师、黄杰老师、贾刚勇老师在我的研究生生涯中给予我的关心和帮助。

感谢东区实验室的兄弟姐妹们，你们让我收获了友情，三年的相处让我们每个人都更加珍惜这份缘分，感谢你们的陪伴。

感谢12届软件所的各位同学，感谢37号楼602的兄弟们，特别要感谢陪我度过三年研究生生活的同寝室兄弟周建成，你每天早上准时七点起床让我记忆犹新，你良好的生活习惯和积极的生活态度是我学习的榜样，感谢你在生活上学习上等各方面对我的支持和鼓励。

感谢生我养我的父母亲，你们为我付出了很多，苍老了很多，接下来也该由

我来回报你们了！还要感谢我的姐姐，你为我的身体健康担忧很多，也希望你能注意自己的健康，工作顺利。

最后，再次向以上的各位老师、同学和朋友表示由衷的感谢，并向百忙之中抽出宝贵时间评审本文的专家、学者和教授们致以最真挚的谢意！

参考文献

- [1] Yin X, Han J, Yang J, et al. Efficient classification across multiple database relations: A crossmine approach[J]. Knowledge and Data Engineering, IEEE Transactions on, 2006, 18(6): 770-783.
- [2] Kuncheva L I, Rodriguez J J. Classifier ensembles with a random linear oracle[J]. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(4): 500-508.
- [3] Baralis E, Chiusano S, Garza P. A lazy approach to associative classification[J]. Knowledge and Data Engineering, IEEE Transactions on, 2008, 20(2): 156-171.
- [4] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions[C]. ICML. 2003: 912-919.
- [5] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine learning, 2000, 39(2): 103-134.
- [6] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]. The eleventh annual conference on Computational learning theory. ACM, 1998: 92-100.
- [7] Joachims T. Transductive inference for text classification using support vector machines[C]. ICML, 1999: 200-209.
- [8] Zhu X, Wu X. Class noise handling for effective cost-sensitive learning by cost-guided iterative classification filtering[J]. Knowledge and Data Engineering, IEEE Transactions on, 2006, 18(10): 1435-1440.
- [9] Yang Q, Ling C, Chai X, et al. Test-cost sensitive classification on data with missing values[J]. Knowledge and Data Engineering, IEEE Transactions on, 2006, 18(5): 626-638.
- [10] 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.
- [11] Abel F, Herder E, Houben G J, et al. Cross-system user modeling and personalization on the social web[J]. User Modeling and User-Adapted Interaction, 2013, 23(2): 169-209.
- [12] 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.
- [13] Sahebi S, Brusilovsky P. Cross-Domain Collaborative Recommendation in a Cold-Start Context: The Impact of User Profile Size on the Quality of Recommendation[M]. User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, 2013: 289-295.
- [14] Li B, Yang Q, Xue X. Can Movies and Books Collaborate? Cross-Domain Collaborative Filtering for Sparsity Reduction[C]. IJCAI. 2009: 2052-2057.
- [15] Pan W, Xiang E W, Liu N N, et al. Transfer Learning in Collaborative Filtering for Sparsity Reduction[C]. AAAI. 2010: 230-235.

- [16] Szomszor M, Alani H, Cantador I, et al. Semantic modelling of user interests based on cross-folksonomy analysis[M]. Springer Berlin Heidelberg, 2008: 632-648.
- [17] Cremonesi P, Tripodi A, Turrin R. Cross-domain recommender systems[C]. Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. IEEE, 2011: 496-503.
- [18] Zhang Y, Cao B, Yeung D Y. Multi-domain collaborative filtering[J]. arXiv preprint arXiv, 2012, 14(3): 89-119.
- [19] Li B, Yang Q, Xue X. Transfer learning for collaborative filtering via a rating-matrix generative model[C]. The 26th Annual International Conference on Machine Learning. ACM, 2009: 617-624.
- [20] Fernández-Tobías I, Cantador I, Kaminskis M, et al. Cross-domain recommender systems: A survey of the state of the art[C]. The 2nd Spanish Conference on Information Retrieval. CERI. 2012: 215-253.
- [21] Li B. Cross-domain collaborative filtering: A brief survey[C]. Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on. IEEE, 2011: 1085-1086.
- [22] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]. The 1994 ACM conference on Computer supported cooperative work, 1994: 175-186.
- [23] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]. The 10th international conference on World Wide Web. ACM, 2001: 285-295.
- [24] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008, 14(1): 1-37.
- [25] Yang Q, Wu X. 10 challenging problems in data mining research[J]. International Journal of Information Technology & Decision Making, 2006, 5(04): 597-604.
- [26] Fung G P C, Yu J X, Lu H, et al. Text classification without negative examples revisit[J]. Knowledge and Data Engineering, IEEE Transactions on, 2006, 18(1): 6-20.
- [27] Al-Mubaid H, Umair S A. A new text categorization technique using distributional clustering and learning logic[J]. Knowledge and Data Engineering, IEEE Transactions on, 2006, 18(9): 1156-1165.
- [28] Sarinnapakorn K, Kubat M. Combining subclassifiers in text categorization: A dst-based solution and a case study[J]. Knowledge and Data Engineering, IEEE Transactions on, 2007, 19(12): 1638-1651.
- [29] Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning[C]. The 24th international conference on Machine learning. ACM, 2007: 193-200.
- [30] Raina R, Battle A, Lee H, et al. Self-taught learning: transfer learning from unlabeled data[C]. The 24th international conference on Machine learning. ACM, 2007: 759-766.
- [31] Daumé III H, Marcu D. Domain Adaptation for Statistical Classifiers[J]. J. Artif. Intell. Res.(JAIR), 2006, 26: 101-126.
- [32] Zadrozny B. Learning and evaluating classifiers under sample selection bias[C].

- The twenty-first international conference on Machine learning. ACM, 2004: 114.
- [33] Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function[J]. Journal of statistical planning and inference, 2000, 90(2): 227-244.
- [34] Dai W, Yang Q, Xue G R, et al. Self-taught clustering[C]. The 25th international conference on Machine learning. ACM, 2008: 200-207.
- [35] Wang Z, Song Y, Zhang C. Transferred dimensionality reduction[M]. Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, 2008: 550-565.
- [36] Dai W, Xue G R, Yang Q, et al. Transferring naive bayes classifiers for text classification[C]. The national conference on artificial intelligence. 2007: 540-556.
- [37] Quionero-Candela J, Sugiyama M, Schwaighofer A, et al. Dataset shift in machine learning[M]. The MIT Press, 2009: 174-176.
- [38] Jiang J, Zhai C X. Instance weighting for domain adaptation in NLP[C]. ACL. 2007, 7: 264-271.
- [39] Liao X, Xue Y, Carin L. Logistic regression with an auxiliary data source[C]. The 22nd international conference on Machine learning. ACM, 2005: 505-512.
- [40] Huang J, Gretton A, Borgwardt K M, et al. Correcting sample selection bias by unlabeled data[C]. Advances in neural information processing systems. 2006: 601-608.
- [41] Bickel S, Brückner M, Scheffer T. Discriminative learning for differing training and test distributions[C]. The 24th international conference on Machine learning. ACM, 2007: 81-88.
- [42] Sugiyama M, Nakajima S, Kashima H, et al. Direct importance estimation with model selection and its application to covariate shift adaptation[C]. Advances in neural information processing systems. 2008: 1433-1440.
- [43] Fan W, Davidson I, Zadrozny B, et al. An improved categorization of classifier's sensitivity on sample selection bias[C]. Data Mining, Fifth IEEE International Conference on. IEEE, 2005: 4-8.
- [44] Dai W, Xue G R, Yang Q, et al. Co-clustering based classification for out-of-domain documents[C]. The 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007: 210-219.
- [45] Ando R K, Zhang T. A high-performance semi-supervised learning method for text chunking[C]. The 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005: 1-9.
- [46] Blitzer J, McDonald R, Pereira F. Domain adaptation with structural correspondence learning[C]. The 2006 conference on empirical methods in natural language processing. Association for Computational Linguistics, 2006: 120-128.
- [47] Lawrence N D, Platt J C. Learning to learn with the informative vector machine[C]. The twenty-first international conference on Machine learning. ACM, 2004: 65.
- [48] Bonilla E, Chai K M, Williams C. Multi-task Gaussian process prediction[J]. 2008, 90(2): 312-330.
- [49] Schwaighofer A, Tresp V, Yu K. Learning Gaussian process kernels via

- hierarchical Bayes[C]. Advances in Neural Information Processing Systems. 2004: 1209-1216.
- [50] Mihalkova L, Huynh T, Mooney R J. Mapping and revising Markov logic networks for transfer learning[C]. AAAI. 2007: 608-614.
- [51] Mihalkova L, Mooney R J. Transfer learning by mapping with minimal target data[C]. The AAAI-08 workshop on transfer learning for complex tasks. 2008: 78-90.
- [52] Davis J, Domingos P. Deep transfer via second-order Markov logic[C]. The 26th annual international conference on machine learning. ACM, 2009: 217-224.
- [53] Loizou A. How to recommend music to film buffs: enabling the provision of recommendations from multiple domains[D]. University of Southampton, 2009: 31-38.
- [54] González G, López B, de la Rosa J L. A multi-agent smart user model for cross-domain recommender systems[J]. Proceedings of Beyond Personalization, 2005, 35(3): 223-244.
- [55] Szomszor M, Alani H, Cantador I, et al. Semantic modelling of user interests based on cross-folksonomy analysis[M]. Springer Berlin Heidelberg, 2008: 8-11.
- [56] AZAK M. CrosSing: A framework to develop knowledge-based recommenders in cross domains[D]. MIDDLE EAST TECHNICAL UNIVERSITY, 2010: 23-26.
- [57] Shi Y, Larson M, Hanjalic A. Tags as bridges between domains: Improving recommendation with tag-induced cross-domain collaborative filtering[M]. User Modeling, Adaption and Personalization. Springer Berlin Heidelberg, 2011: 305-316.
- [58] Singh A P, Gordon G J. Relational learning via collective matrix factorization[C]. The 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 650-658.
- [59] Hu Y, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets[C]. Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008: 263-272.
- [60] Ding C, Li T, Peng W, et al. Orthogonal nonnegative matrix t-factorizations for clustering[C]. The 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 126-135.
- [61] Lee D D, Seung H S. Algorithms for non-negative matrix factorization[C]. Advances in neural information processing systems. 2001: 556-562.
- [62] Gao S, Denoyer L, Gallinari P. Temporal link prediction by integrating content and structure information[C]. The 20th ACM international conference on Information and knowledge management. ACM, 2011: 1169-1174.
- [63] Ding C, Li T, Jordan M I. Convex and semi-nonnegative matrix factorizations[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(1): 45-55.

附录：

作者在读期间发表的学术论文及参加的科研项目

发表的学术论文

1. Jian Wan, Xin Wang, Yuyu Yin, Renjie Zhou. Cross-domain Recommendation combining Feature Tags with Transfer Learning[J]. International Journal of Grid and Distributed Computing. (Accepted, EI journal).

参加的科研项目

1. 国家科技支撑计划：支撑智慧城市的海量数据共享与智能分析应用示范 (2012BAH24B04), 2012-2014.
2. 浙江省科技计划重大专项：基于存储集群的内容智能管理系统研究及产业化 (2011C11038), 2011-2013.
3. 教育部重大专项：中国教育科研网格 ChinaGrid 二期——力学学科应用网格，2012-2013.