密级：_____

浙江大学

硕 士 学 位 论 文

论文题目 基于隐式因子和隐式主题的跨域推荐算法研究

作者姓名 肖力涛

指导教师 陈德人 教授

郑小林 副教授

学科(专业) 计算机应用技术

所在学院 计算机科学与技术学院

提交日期 2016 年 1 月

A Dissertation Submitted to Zhejiang University for the Degree of Master of Engineering

TITLE: Research of Cross-domain Recommendation System based on Hidden Factors and Hidden Topics

Author: Litao Xiao

Supervisor: Prof. Deren Chen

Asso. Prof. Xiaolin Zheng

Subject: Computer Applications Technology

College: Computer Science and Technology

Submitted Date: 2016-01

浙江大学硕士学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，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 **浙江大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：　　签字日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解 **浙江大学** 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 **浙江大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　导师签名：

签字日期：　　年　　月　　日　　签字日期：　　年　　月　　日

摘要

随着互联网的发展，推荐系统作为一种个性化的个人信息过滤系统变得日益重要，使得用户能够在海量数据中，迅速获取自己需要的信息。现有的推荐模型更多建立在单一域的用户信息之上，利用矩阵分解得到潜在向量，之后做出预测。然而用户不同域之间的数据往往不是独立的，并且在利用主题模型对用户评论建模时，更多的是简单将主题因子加入到潜在向量中。

引入跨域思想（Cross-domain），并充分利用用户的评论信息，本文提出了基于隐式主题和隐式因子的跨域推荐模型（Cross-domain Recommendation Model based on Hidden Factors and Hidden Topics）。首先，我们利用主题模型建立潜在向量与主题因子之间的映射关系，将用户的评论信息作为反馈引入到传统的概率矩阵分解模型当中，完成对单一域之上的建模；之后，我们通过建立不同域之间潜在用户向量的非线性映射，将不同域结合起来，形成跨域；最终，利用训练得到的潜在向量进行分数预测。

论文在国内知名的推荐领域网站豆瓣网的数据集上进行结果分析和对比实验。实验结果表明，相较于传统的单一域推荐模型，本文模型在相对较为稀疏的数据集上，预测效果有了明显的提升。

关键词：推荐系统，跨域，主题模型，矩阵分解，协同过滤

Abstract

With the growth of internet, people pay more attention to recommendation system, which make users quickly get useful information. Most existing recommendation model focus on single domain, while ignoring the review texts and the relation between different domains.

In this paper, we investigate how to utilize the review texts and cross-domain information. Given this, we propose Cross-domain Recommendation Model based on Hidden Factors and Hidden Topics. First we combine latent rating dimensions (such as those of latent-factor recommender system) with latent review topics (those learned by topic models). And then we build a non-linear mapping functions between latent user factors in different domain. At last, we use the final latent factors to predict the users' rating.

The experiments are based on the data from douban which is well-known recommendation website. Experimental verifications have demonstrated, our model have a considerable improvement over the previous single-domain model.

Keywords： recommendation system, cross-domain, topic model, matrix factorization, collaborative filtering

目录

图目录

表目录

第1章 绪论

1.1 课题背景

随着互联网的发展，网络所蕴含的资源信息越来越丰富，然而如此丰富的在线资源也导致了过度膨胀的信息量。例如，如果一个用户想要购买一款数码相机，那么让用户自己去收集网络上的所有相关信息然后再做出购买决定，这几乎是一项不可能完成的任务。而推荐系统通过向用户推荐一系列他可能感兴趣的商品来解决类似的问题。精准的推荐使得用户能够快速准确的定位到他感兴趣的商品，从而极大的将无关信息过滤掉，提高用户的购买效率[1]。同时，对于销售商而言，推荐系统所带来的好处是不言而喻的，精准的推荐可以极大的提高商品的销售量。

个性化推荐即根据用户的购买行为，分析用户的兴趣点，推荐用户感兴趣的商品，一方面降低用户的搜索成本，另一方面提高商用系统的销售量。常见的网站，例如国外的 Amazon[2]、Netflix，国内的如豆瓣（国内知名的评论推荐网站，涉及电影、书籍、音乐等）、淘宝等。这些网站会向用户推荐包括电影、书籍、歌曲、文章、商品等等生活中方方面面的内容，而在推荐的过程当中，推荐的准确性，以及推荐算法的效率等问题就是推荐算法研究的着重点。

推荐系统可追溯到很多相关研究领域，例如认知科学、近似理论和信息检索等等。由于其与日俱增的重要性，它在 1990 年左右发展成一个独立的研究领域。简单来说，推荐系统的任务无非是向用户推荐其感兴趣的商品。从推荐方法上来讲，大致可以分为：基于内容的推荐算法[3]以及基于协同过滤的推荐算法。协同过滤算法又可以被分为基于近邻的协同过滤、基于模型的协同过滤以及混合协同过滤[4, 5]。

基于内容的推荐算法主要思想是根据用户过往的行为记录，通过建立模型挖掘出用户的特征向量和物品的特征向量，将相似特征的物品推荐给用户。这种方

式的好处很明显，透明度高，推荐方式直接，而且当有新物品出现时，加入新的特征值即可进行推荐。然而缺点较为明显，物品的内容特征会很难提取，因为内容有时可能包括图片、视频文件、文本等等，难度较大。

由此研究者提出了基于协同过滤的推荐算法[6]，而该算法也是目前推荐系统研究的热点之一，大多数推荐算法都是在此基础上改进而来。协同过滤算法的核心是围绕用户-物品评分矩阵展开的，即每个物品，当用户使用过之后给出一个评分，例如我们常见的豆瓣电影评分，用户看过电影之后，会给出 5 分制的一个评分标准。这样就会形成一个用户和物品的评分矩阵。这样利用评分矩阵，最基本的两种协同过滤算法便是基于近邻-物品的协同过滤算法以及基于近邻-用户的协同过滤算法。基于近邻-用户的协同过滤算法是从用户的角度来使用评分矩阵，例如用户 A 和用户 B 都看过电影 m1 和 m2，并且都认为 m1 是烂片而 m2 是好片，那么我们可以假设用户 A 和 B 拥有相同的兴趣爱好，那么如果用户 B 还看过电影 m3，我们便可以将电影 m3 推荐给用 A。而类似的从物品的角度利用评分矩阵进行推荐，同理发现物品的相似度，然后根据目标用户的历史偏好信息，将物品推荐给用户。而在度量用户与用户或者物品与物品之间的相似度的时候，可以利用欧几里得相似度、余弦相似度、皮尔森相似度等等。这种最基本的协同过滤算法存在着很明显的缺陷，首先用户评分矩阵往往是非常稀疏的，所以在度量相似度时会存在问题；其次当一个新的用户进入系统时，由于没有任何新用户的历史信息，往往无法进行有效的推荐，即冷启动问题；而且，随着用户以及物品数量的不断增加，用户向量以及物品向量的维度会急剧提升，这样会导致算法的运算效率急剧下降。为了解决这些问题，人们在原有协同过滤的算法上不断改进，提出了新的算法。

基于模型的协同过滤算法，其中具有代表性的是奇异值分解 SVD（singular value decomposition）[7]和概率矩阵分解 PMF（probability matrix factorization）[8]，这两种方法的共同特点是对原有的评分矩阵进行了分解，而不是简单的在原始数据上进行操作。这样从一定程度上解决了矩阵的稀疏性问题。而在此基础上，研

究人员又将用户的社交关系引入到推荐算法当中，利用用户之间的社交关系，进一步提高推荐的准确率。另一方面，研究人员通过主题模型，将用户的评论信息引入到推荐系统当中，提高了准确率。

基于上面的阐述，我们会发现现有模型更多的是针对单一域的用户信息进行建模，而且对于用户的评论信息重视程度不够。如果能够充分利用不同域间的用户信息，并且更加有效的利用用户的评论信息，能够在一定程度上解决矩阵稀疏和冷启动问题，而本文基于此，展开将跨域思想和用户评论的主题建模引入到传统的推荐模型中的研究，以期达到更好的推荐效果。

1.2 国内外研究现状

本节对推荐系统目前的研究现状做一个介绍，从最基本的基于关联的推荐、基于内容的推荐，到基于协同过滤的推荐，目前多数推荐算法都是在基于协同过滤的推荐算法上改进而来。

1.2.1 基于关联规则的推荐

基于关联的推荐是以关联规则为基础[9, 10, 11, 12]，把已购商品作为规则头，推荐对象作为规则体，从大量数据中挖掘出不同商品在销售过程中的潜在联系。最为经典的案例是发生在美国沃尔玛超市中的一个案例：零售超市将两个看似毫无关联的商品啤酒与尿布放在了一起，发现两者的销量都得到显著提高，原来美国妇女一般在家照顾孩子，所以她们会嘱咐下班的丈夫给孩子带尿布，而丈夫一般还会顺手为自己带一瓶啤酒，这个发现为商家带来了大量的利润。所以这种推荐也成了最早期的一种推荐方式，相关研究如 Apriori 算法和 F-P 算法就是用来大量数据中发掘出这种关联规则又称频繁项[13, 14]。

这种推荐方式的缺点也很明显，首先是算法的计算复杂度较高，并且商品的同义性等原因也会导致推荐准确率不高。

1.2.2 基于内容的推荐

用看过的内容作为检索关键字

基于内容的推荐方法的理论依据主要来源于信息检索和信息过滤，该方法利用用户过去的行为记录，利用 *tf-idf*（term frequency–inverse document frequency）[15]或者其他模型方法来抽象出描述用户的特征向量，之后用同样的方法抽象出推荐项的特征向量，最后将与用户特征向量相似的物品推荐给用户，而相似性的计算一般会使用 *cosine* 方法。图 1-1 描述了基于内容推荐的基本原理：

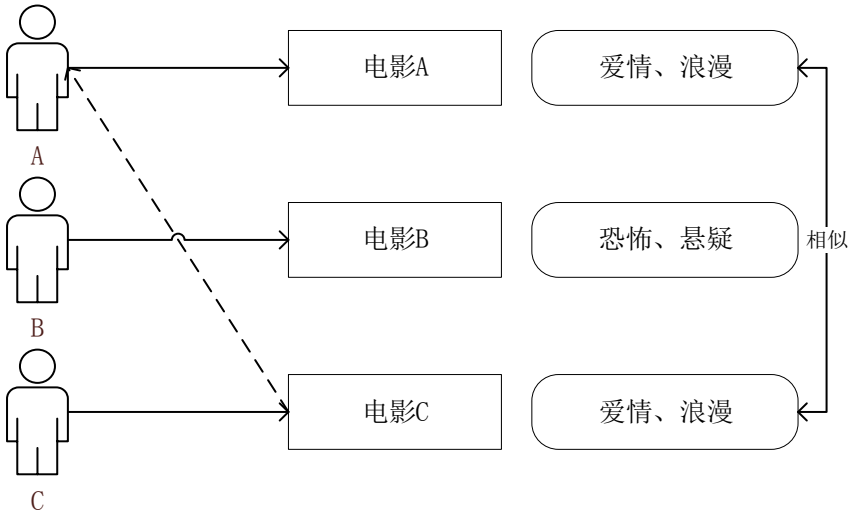

图 1-1 基于内容的推荐原理

用户 A 喜欢电影 A，我们从中提取出用户 A 对爱情以及浪漫等特征感兴趣，之后我们对其他推荐项进行分析，可以发现电影 C 符合这种特征，那么我们将电影 C 推荐给用户 A。

基于内容的推荐很好的利用了用户的过往行为，充分挖掘用户的喜好，但也存在一些问题：

1. 需要对物品进行分析和建模，推荐的质量受到物品模型完整性和全面性的影响，自动化程度较低
2. 只能推荐和用户兴趣相匹配的物品，无法发掘用户新的兴趣点
3. 对于音乐和艺术品等难以用特征描述的物品，无法进行建模，推荐难度

1.2.3 基于协同过滤的推荐

1992 年 David Goldberg[16]提出一个实验性的邮件系统 Tapestry，该邮件系统利用小型社区成员的观点对邮件进行过滤，而该系统也是已知的最早的协同过滤推荐系统。之后随着互联网的不断发展，尤其是电子商务的出现，推荐算法也随之快速发展，而目前大部分算法都是基于协同过滤算法改进而来。

1.2.3.1 基于近邻的协同过滤

基于近邻的协同过滤方法是最基本的协同过滤方法，将用户物品评分矩阵作为输入，通过计算向量之间的距离来度量物品与物品或者用户与用户之间的相似度，找出最相似的对新物品的评分进行预测。

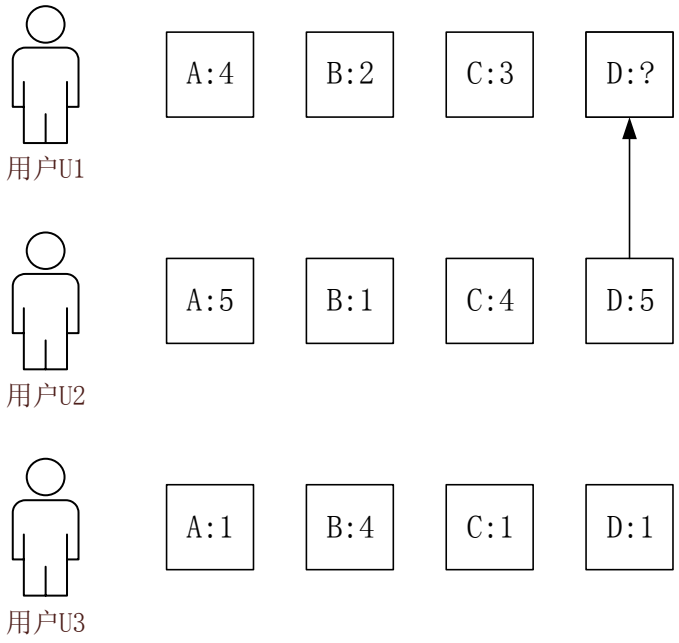

图 1-2 基于近邻-用户的协同过滤方法

我们以基于近邻-用户的协同过滤方法为例（另一种方法类似），如图 1-2 中，假设系统中现有三个用户 U1，U2，U3 以及物品 A、B、C、D，用户以 5 分制对物品进行打分，图中我们可以看出三个用户的评分向量分别为$U_1 = [4, 2, 3, ?], U_2 = [5, 1, 4, 5], U_3 = [1, 4, 1, 1]$，我们发现用户 U1 对物品 D 的评分是

未知的，所以我们不知道用户 U1 对物品 D 的喜好程度。通过计算我们发现，用户 U1 与用户 U2 的评分向量相似度较高，那么我们将用户 U2 关于物品 D 的评分预测给用户 U1。当然在度量用户向量间的距离的时候，我们可以采用不同的度量方法，如余弦相似度、皮尔森相似度等等 [6]。

基于近邻的协同过滤算法在用户评分矩阵并不稀疏的时候能够产生非常良好的效果，而且该算法并不需要训练，直接通过计算就可以给出推荐，但是当数据量非常庞大的时候，推荐过程往往伴随着大量的计算，这也从一定程度上阻碍该算法在线上系统当中的使用。

1.2.3.2 基于矩阵分解协同过滤

将用户-物品高维度的评分矩阵分解为低维度的特征矩阵是协同过滤中最成功的算法之一，这种模型的基本假设是一个用户的喜好或者观点往往取决于很小一部分潜在特征。因此在一个线性的特征模型中，一个用户的喜好是线性的将特征向量结合来建模的。例如，对于 N 个用户和 M 部电影而言，其$N \times M$的评分矩阵 R 是由$N \times D$的用户潜在特征矩阵U^T和一个$D \times M$的物品潜在特征矩阵 V 相乘得到（D ≤ M & D ≤ M）。训练该模型就是在给出的损失函数下，找到最优的秩 D 使得$U^T V$最接近于已观察到的$N \times M$的目标矩阵 R。

基于以上的描述，我们可以得出矩阵分解模型的基本流程如下图 1-3。

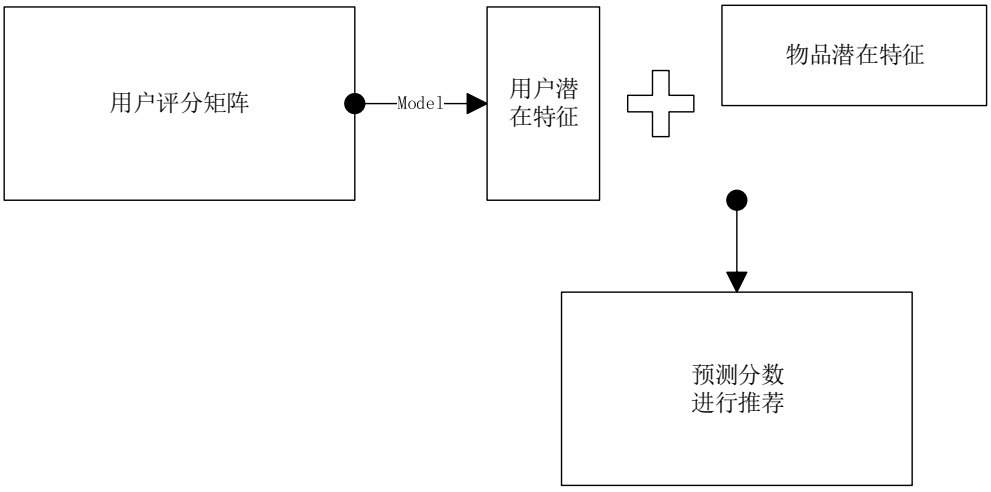

图 1-3 基于模型的协同过滤方法

其中最基本的矩阵分解是使用奇异值分解 SVD（Singular Value Decomposition）[17]。SVD 在给定矩阵的秩的情况下，计算得到矩阵$\hat{R} = U^T V$，使得与目标矩阵 R 的平方和误差最小。但是大多数真实的数据集都是非常稀疏的，所以在 R 中的很多数据都会缺失，从而使得平方和误差仅仅由目标矩阵的少数被观察点来计算，这就导致 SVD 算法的效果非常的差。

为了克服这些缺点，Ruslan 和 Andriy 便提出了概率矩阵分解的方法 PMF（Probabilistic Matrix Factorization）[8]，PMF 将高斯分布引入到矩阵分解当中，并通过贝叶斯公式得出损失函数以求得特征因子矩阵 U 和 V，并最终做出预测。

基于矩阵分解的算法在推荐系统中有较好的表现，而且从一定程度上解决了用户评分矩阵的稀疏性问题，后续的很多学者也在矩阵分解的基础上不断扩展深入，如下文会提到的将社交以及主题模型等算法引入进来，从而达到更好的推荐效果。

1.2.3.3 基于社交网络的推荐

在线社交网络的发展，为提高推荐准确率提供了新机遇。在现实生活中，人们在进行购买商品或者消费服务的时候往往会受到其社交网络中朋友的影响，而

且社会学以及心理学研究也表明，人们趋向于联合与团结相似的人。所以，人们通常更愿意相信自己熟悉的人的推荐而非陌生人或者销售商。显然我们可以将社交网络结合到推荐系统中，可称为社交推荐[18]。

通过将社交网络中用户间的社交关系和社交信任作为额外的输入可以提高推荐的准确率。例如，由于社交关系，一个用户可能对某篇新闻报道的发生在某地的事件非常感兴趣，这仅仅是因为他的家住在那里；再如，由于社交信任，一位用户可能会喜欢来自其最好的朋友推荐的一本书。在社交推荐的研究方面，Massa[19, 20]首先提出了将信任关系融入到推荐系统当中，在基于用户的协同过滤算法中，将信任结合到相似度中，提出了“信任权重”的概念作为相似度的结果。之后，Ma[21]提出了 STE（Social Trust Ensemble）的方法，将信任列表中用户的口味结合到用户自己的评分当中，以加权的方法对用户评分矩阵进行分解。在 STE 模型的基础上，Jamili[22]利用社交信任关系约束潜在因子进行矩阵分解，提高了推荐效果。Ma 提出了 SoRec(Social Recommendation)模型，不同于之前研究人员直接将信任关系融入到模型中，Ma 利用用户间的信任关系形成用户信任矩阵，并分别对用户物品评分矩阵以及用户信任矩阵进行概率矩阵分解，再将两者结合。Chen[23]结合了 STE 和 SoRec 的优点，提出了混合的 RSTR（Recommend on Social Trust Relationships）模型。在社交关系上面进一步挖掘，Yang[24]提出了“社交圈”的概念，并且认为用户对不同的社交圈的信任程度是不同的，从而提供不同的权重，并将这种关系应用到推荐当中。与此类似的做法，Tang[25]则认为用户对不同领域的信任程度不同，从而将这种关系加入到推荐中。

上面我们多提到的是一种信任关系，我们可以发现信任关系往往是单向的，即用户 A 信任用户 B，但是反过来，用户 B 不一定信任用户 A。所以也有学者将好友关系应用到推荐中，好友关系不同于信任关系，它是一种双向的关系。Ma[26]提出了在将好友关系替换原有的信任关系应用到社交推荐中，从而取得了更好的效果。同样的 Chen[27]在研究过程当中也利用了好友关系对潜在因子进行约束。

在评分矩阵之外，由于将社交反馈信息加入到推荐系统当中，使得推荐的效果有了一定的提升。但是在加入用户社交关系的同时，也使得模型的计算成本更加的复杂，同时社交关系的模型往往也只是针对单一域内，将用户的社交关系作为约束引入进来。

1.2.3.4 基于主题模型的推荐算法

Blei[28, 29]提出了主题模型 LDA（Latent Dirichlet Allocation），这是一个经典的统计模型，最初用于文本语义挖掘，用来发现文档中隐含的主题，将词项空间表达的文本映射到低维的主题空间，从而实现对文档的分类以及检索等。在推荐系统的研究过程当中，许多研究人员将 LDA 引入进来，对推荐场景中的文本或者其他信息进行建模，并将这些信息作为约束或者反馈信息加入到矩阵分解模型当中。Chong Wang[30]提出了协同主题回归模型（CTR , Collaborative Topic Regression），用于对科技文献的推荐，利用物品的主题分布向量约束物品的潜在因子，结合到概率矩阵分解模型中，从而产生更好的效果。在此基础上，研究人员将社交关系加入到 CTR 主题模型当中，其中具有代表性的是 Purushotham[31]提出的 CTR-SMF（Collaborative Topic Regression with Social Matrix Factorization）模型，将社交信任矩阵引入到模型当中。Kang[32]提出了关注度的概念，将用户在物品上的感知信息引入的 CTR 模型中。另外一个对于 CTR 模型的改进是 Julian[33]提出的隐式因子结合隐式主题模型（HFT, Hidden Factors and Hidden Topics）模型，不同于 CTR 模型将 LDA 建模得到的反馈信息直接加到矩阵分解的物品的潜在向量中，HFT 模型认为主题因子的先验并不是关联于外在的变量，而是与潜在因子（潜在用户向量和潜在物品向量）有关，所以模型建立了一种物品潜在向量到 LDA 中主题分布向量 θ 的映射，从而提高了模型的整体效果。

将主题模型结合到推荐系统中，是当今推荐系统的研究热点之一，很多推荐算法模型都是建立在主题模型的基础之上，本文所提出的模型也会将主题模型融入进来。

1.2.3.5 基于跨域的推荐模型

在前文我们所叙述的模型中，我们会发现研究人员在建立模型的过程当中往往是针对单一域而言的，即如果我们给用户推荐电影，则只考虑用户对电影的评分记录；同样给用户推荐音乐，则只考虑用户对音乐的评分记录。因此是在单一域上的建模，那么如果将两者结合起来推荐，例如喜欢摇滚音乐的用户我们可以给他推荐科幻片，喜欢抒情音乐的用户我们可以给她推荐爱情片（一种直观的假设），这样就提出了推荐系统当下研究热点之一，跨域推荐。

Leizou[34]提出了跨域推荐的三个主要研究趋势：

- 将不同域中的用户特征集成到一个统一的域当中
- 通过互联网监测用户的行为来描述用户的特征
- 为了提高一个目标域的推荐效果，将单一域集成起来

对于第一个趋势，Gonzalez[35]提出在每一个域上建立用户模型，之后通过域之间的关系将它们统一到被称作 smart 的用户模型中。但是文章并没有具体设计所谓 smart 域的一些细节。

而对于第二个趋势，Tuffield[36]建议使用被称为语义日志（Semantic Logger）的一种元数据获取引擎，用来无监督的抓取任何由用户产生或使用过的信息（来自邮件、日历记录、URLs、标签等）。这些琐碎的信息通过一个共同的词表和相互比较被整理到一个共同的空间。数据源之间的关系依靠一种面向图的方式，而最终推荐是利用一种基于马尔科夫链的基于图的算法，但是该算法在大型数据量上几乎不可计算。Lee[37]提出从用户的历史行为中提取关联规则，然后将用在目标用户的未知评分物品中。关联规则是基于近邻用户进行学习，但是同样在大型数据集上变得不可计算。

最后，对于单一域推荐系统的集成，Berkovsky[38]等提出将多资源整合到目标域的四种集成方式：（1）标准（standard），外在的资源数据集成到目标域来丰富

评分数据；（2）启发式（heuristic），与用户评价相似的物品列表会在源数据域和目标域之间共享；（3）跨域（cross-domain），相似的物品列表会在两者之间共享由源数据域计算出的相似度；（4）远程（remote-average），目标域会集成由源数据集所推荐的物品。

Zhuang 等[39]在进行跨域建模的时候，通过香农熵和一种逻辑回归来进行一致性计算，从而在不同域之间建立关联。Cao[40]等使用一种相似的方法解决称为连接预测的问题，即预测在用户和物品之间的潜在连接，其提出一种概率贝叶斯框架来进行这种计算。

跨域推荐结合到传统的经典推荐算法中是当今推荐系统研究的热点之一，不同域之间数据的共享丰富了数据信息，从而能够大大提高推荐的准确率。同时怎么组织不同域之间的数据，以及与推荐算法的结合也是跨域推荐的难点。本文提出的模型会充分利用跨域，将跨域的思想融合到我们自己的模型当中。

1.3 本文的主要工作

本文的研究内容是提出一种基于隐式因子和隐式主题的跨域推荐算法（Cross-Domain Recommendation System based on Hidden Factors and Hidden Topics，简称 Cross-HFT）。上文我们提到了很多经典的推荐算法，而这些推荐算法往往是针对单一域而言进行建模。本文在 PMF 推荐模型的基础上，通过建立潜在向量与主题因子之间的映射，对用户评论进行建模，并将跨域推荐的思想引入进来，在不同域之间找到一种非线性的映射关系，使得新的模型在较为稀疏的数据集上有了明显的提升。

本文首先利用 HFT 模型在目标域和辅助域上进行建模，之后通过非线性的映射函数，将目标域和辅助域关联起来，形成共同的目标函数，最终我们利用 KKT 条件、梯度下降和吉布斯采样的方法对目标函数进行求解。

本文实验所采用的数据集来源于豆瓣网，数据集包含用户多个域的使用记录，

以及用户的评论信息，我们会利用这些数据对我们的模型进行实验与测试，并做出详细的统计与分析。

1.4 本文组织结构

本文总共分为五个章节：

第一章是绪论，主要介绍了本文的课题背景，以及国内外的研究现状，并针对传统的模型的问题进行分析，并提出本文的主要内容和研究思路。

第二章是相关技术综述，从矩阵分解模型入手，之后详细介绍了主题模型的思想，主题模型结合概率矩阵分解进行推荐，以及跨域推荐的思想。

第三章阐述了本文所提出的利用评论信息的跨域隐式因子和隐式主题推荐模型，从模型的定义以及模型的框架和设计流程分别进行了阐述。

第四章是实验与分析，从数据集的来源，评价标准以及参数的训练等方面进行了介绍，并对实验结果进行分析，将 Cross-HFT 与 CF 算法、PMF 算法、HFT 算法、Cross-CTR 模型进行比较，论证本文提出模型的有效性。

第五章对课题研究进行了总结与展望，总结了全文的研究内容，并对未来的改进方向做出了展望。

第2章 相关技术综述

2.1 协同过滤算法

在前文我们已经提到协同过滤算法，在这一节对一些细节进行阐述。协同过滤算法不同于基于内容的推荐算法，它利用用户的历史信息，建立特定的模型，从中挖掘出用户兴趣并做出推荐。本节主要阐述两种最主要也是最基本的协同过滤算法：基于近邻的协同过滤和基于矩阵分解的协同过滤。

2.1.1 基于近邻的协同过滤算法

2.1.1.1 基于用户的近邻协同过滤

在前文我们阐述了基于近邻协同过滤模型[6]的基本思想，即如果通过模型我们发现用户 A 和用户 B 具有相似的兴趣爱好，那么我们就可以将用户 A 好评的物品推荐给用户 B，同样用户 B 好评的物品推荐给用户 A。这是非常直观的一种思想，那么如何判断或说度量两个用户之间的相似度，也是近邻协同过滤的关键所在。

假设用户-物品评分矩阵为 R 是一个$n \times m$维的矩阵（n 个用户，m 个物品），我们将用户 i 对物品的评分表示成一个 m 维向量$u_i = (r_{i1}, r_{i2}, \dots, r_{im})$，其中$r_{ij}$表示用户 i 对物品 j 的评分，在近邻模型中一般通过计算评分向量之间的相似度从而获得用户之间的相似度。常用的相似度计算方式有：欧几里得相似度、余弦相似度、皮尔森相似度等。

为了更好的描述相似度的度量方法，我们假设用户$u_i = (r_{i1}, r_{i2}, \dots, r_{im})$和用户$u_j = (r_{j1}, r_{j2}, \dots, r_{jm})$，下面我们分别阐述几种相似度计算方式：

（1） 欧氏距离：

$$d(i, j) = \sqrt{\sum_{K \in I_{ij}} \left(r_{ik} - r_{jk}\right)^2} \tag{2.1}$$

其中I_{ij}表示用户 i 和用户 j 共同评分过的物品，r_{ik}和r_{jk}分别表示用户 i 和用户 j 对物品 k 的评分。其中欧氏距离的值域为 0 到无穷大，为使得相似度规约到（0,1）之间，经过变形我们可以得到欧式相似度公式：

$$Sim(i,j)=\frac{1}{d(i,j)+1}=\frac{1}{\sqrt{\sum_{k\in I_{ij}}\left(r_{ik}-r_{jk}\right)^2}+1} \tag{2.2}$$

（2） 余弦相似度

$$Sim(i,j)=\cos\left(\vec{i},\vec{j}\right)=\frac{\vec{i}\cdot\vec{j}}{\left\|\vec{i}\right\|_2*\left\|\vec{j}\right\|_2} \tag{2.3}$$

其中·代表两个向量的点积，余弦相似度即是两个向量间夹角的余弦值，从中我们可发现其更多刻画的是两个向量方向上的相似度，而对数值的大小并不敏感，比如用户 A 和 B 的评分分别为（1,2）和（4,5），我们计算余弦相似度会发现为 0.98，相似度很高，但是其真实的评价值，用户 A 很明显不喜欢这两个物品而用户 B 是喜欢的，而另外一个常用的相似度便是皮尔森相似度。

（3） 皮尔森相似度

$$Sim(i,j)=\frac{\sum_{k\in I_{ij}}\left(r_{ik}-\overline{r_i}\right)\left(r_{jk}-\overline{r_j}\right)}{\sqrt{\sum_{k\in I_{ij}}\left(r_{ik}-\overline{r_i}\right)^2}\sqrt{\sum_{k\in I_{ij}}\left(r_{jk}-\overline{r_j}\right)^2}} \tag{2.4}$$

其中$\overline{r_i}$和$\overline{r_j}$分别表示用户 i 和 j 所有评分过的物品的平均分数。皮尔森相似度考虑了有些用户评分整体过高或过低的情况，对这部分用户做了平衡。

2.1.1.2 基于物品的近邻协同过滤

与基于用户的近邻协同过滤模型类似，以物品为主，如果用户对于物品 A 和物品 B 的所有评分都相似，那么我们可以假设物品 A 和物品 B 是相似的，那么我们可以将物品 B 推荐给喜欢物品 A 的用户。其关键也是在于相似度的计算方式，这里我们就不再赘述。

2.1.2 矩阵分解模型

基于近邻的协同过滤算法是一个经典的算法，而目前推荐算法研究的热点更多集中在对矩阵分解模型的扩展以及深入，最基本的两个矩阵分解模型分别是低维因子模型和概率矩阵分解。

2.1.2.1 低维因子模型

从直观的例子出发，假设今年的新片《捉妖记》的搞笑因子是 0.8，而它的科幻因子是 0.2。一位用户 A 可能比较喜欢偏喜剧的电影，假设他对搞笑因子的喜欢程度为 0.7，而对科幻因子的喜欢程度为 0.3。那么我们会发现用户 A 和电影《捉妖记》的特征就非常吻合，随之用户 A 对其评分也会非常高。低维因子就是从这种想法出发，去挖掘用户和物品的特征因子。

我们用形式化的语言来描述模型，假设有 N 个用户和 M 个物品，其评分矩阵为 R 为 $N \times M$ 维矩阵，那么我们可以将矩阵 R 分解如下：

$$R = P^T Q \tag{2.5}$$

其中 P^T 为 $N \times K$ 维的低维用户因子矩阵，而 Q 为 $K \times M$ 维的低维物品因子矩阵（$K<N\&\&K<M$）。在实际情况中，由于存在噪音，以及用户评分的喜好（如普遍评分偏低或偏高），因此我们将一些偏移量加入到公式中，将公式改进如下：

$$R = OverallMean + b_U + b_I + P^T Q \tag{2.6}$$

矩阵每一项的计算公式如下：

$$r_{u,i} = \mu + b_u + b_i + p_u^T q_i \tag{2.7}$$

其中 μ 为所有评分的平均分，b_u 和 b_i 分别表示 用户 u 和物品 i 的偏移量，p_u 和 q_i 分别表示用户 u 和物品 i 的特征因子。 与总体的偏差，表示用户的特殊喜好

最后为了训练我们的模型，我们需要最小化预测值和真实值的误差，我们可以得到如下目标函数：

$$\min(\sum_{u,i}(r_{ui}-\mu-b_u-b_i)^2+\lambda(\|p_u\|^2+\|q_i\|^2+b_u^2+b_i^2)) \quad (2.8)$$

目标函数的第一项为误差计算项，而第二项为防止过拟合的项。

2.1.2.2 概率矩阵分解模型

低维因子模型是最基本的矩阵分解方法，但是该方法在计算的过程当中更多的是基于已知的评分数据，而在真实的场景中，往往数据是非常稀疏的，这就导致该模型推荐效果非常差。而基于此 Ruslan[8]等便提出了概率矩阵分解模型（PMF）。

假设我们有 M 部电影和 N 位用户，用户的评分是（1，K）之间的整数。R_{ij} 表示用户 i 对电影 j 的评分，$U\epsilon R^{D\times N}$和$V\epsilon R^{D\times M}$是潜在的用户和电影特征矩阵，其中列向量U_i和V_j分别表示用户特征向量和电影特征向量。由于模型的评价标准是在测试集上采用平方根均值误差（RMSE），所以我们采用带有高斯噪音的线性概率模型。于是我们将已观测数据的条件概率分布定义如公式(2.9)：

$$p(R\,|\,U,V,\sigma^2)=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[N(R_{ij}\,|\,U_i^TV_j,\sigma^2)\right]^{I_{ij}} \quad (2.9)$$

其中$N(x|\mu,\sigma^2)$是高斯分布的概率密度函数，而I_{ij}是一个指示函数：如果用户 i 评价过电影 j 其为 1，否则为 0。同时我们在用户和电影特征向量上还定义了 0 均值的高斯先验：

$$p(U\,|\,\sigma_U^2)=\prod_{i=1}^{N}N(U_i\,|\,0,\sigma_U^2I) \quad (2.10)$$

$$p(V\,|\,\sigma_V^2)=\prod_{i=1}^{M}N(V_j\,|\,0,\sigma_V^2I) \quad (2.11)$$

则用户和电影特征的后验概率为：

$$Inp(U,V \mid R,\sigma^2,\sigma_V^2,\sigma_U^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij} - U_i^T V_j)^2$$
$$-\frac{1}{2\sigma_U^2}\sum_{i=1}^{N} U_i^T U_i - \frac{1}{2\sigma_V^2}\sum_{j=1}^{M} V_j^T V_j \qquad (2.12)$$
$$-\frac{1}{2}\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\right) In\sigma^2 + NDIn\sigma_U^2 + MDIn\sigma_V^2\right) + C$$

其中 C 是不依赖于其它参数的常数。在固定超参数（观测噪音方差和先验方差）的前提下，利用电影和用户特征最大化后验概率，等价于最小化带有二次正则项的平方和误差的目标函数：

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij} - U_i^T V_j)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_{Fro}^2 + \frac{\lambda_V}{2}\sum_{j=1}^{M}\|V_j\|_{Fro}^2 \qquad (2.13)$$

其中$\lambda_U = \sigma^2/\sigma_U^2$，$\lambda_V = \sigma^2/\sigma_V^2$，而$\|\cdot\|_{Fro}^2$表示 Frobenius 正则。PMF 的图模型如图 2-1：

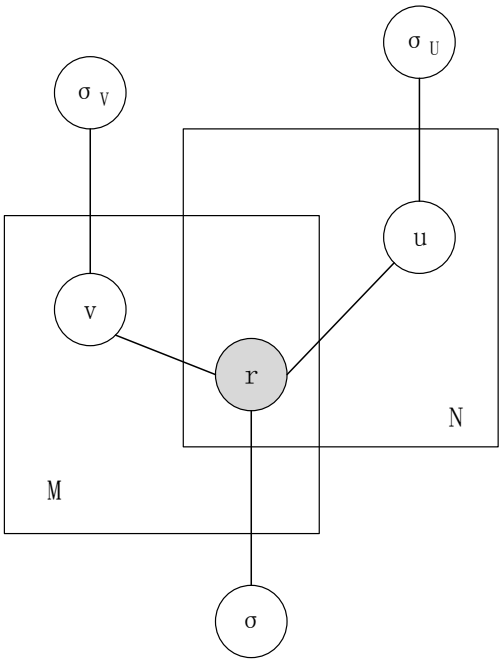

图 2-1 PMF 图模型

PMF 从一定程度上解决了大规模数据出现的矩阵稀疏性问题，但是也存在着一定局限性。而本文要讨论的模型从广义的角度讲也是从 PMF 上扩展而来。

2.2 主题模型相关技术

将主题模型融入到推荐系统当中，是当前研究的热点之一，其中应用最为广泛的便是隐式狄利克雷分布（LDA, Latent Dirichlet Allocation），它是一种主题模型，可以将文档集中每篇文档的主题按照概率分布的形式给出，本节将就主题模型的演化以及一些相关技术做一个简单的阐述。

2.2.1 主题模型的演化

文本挖掘技术一直是研究领域的热门话题，最早利用基于向量空间的模型（Vector Space Model）[41, 42, 43]来对文本进行建模，之后 Landauer 等人进一步提出了潜在语义分析模型（Latent Semantic Analysis, LSA）模型[44]和 Hofmann 将概率引入后的概率潜在语义分析（Probabilistic Latent Semantic Analysis, PLSA）模型[45]，对于 VSM 存在的一词多意和一意多词进行了解决。但是 PLSA 的可扩展性较差，随着文档数量的增加，其计算复杂度会持续上升，导致计算量过大。对此 Blei 等人便提出了经典的 LDA（Latent Dirichlet Allocation）主题模型[29]，该模型在研究领域得到了极为广泛的应用。

2.2.2 模型定义和术语

首先我们介绍 LDA 模型中的术语如“单词”，“文档”，“语料库”。定义如下：

- 一个单词是离散数据中最基本的单元，定义为由$\{1,\ldots,V\}$索引的单词表中的一项。我们用向量中的一个分量来表示单词，有表示 1，无表示 0。
- 一个文档是 N 个单词的一个序列，其基于“词袋”的概念，即文档中单词的顺序无关，表示为$\boldsymbol{W} = (w_1, w_2, \ldots, w_N)$，其中$w_n$表示序列中的第 n 个单词。
- 一个语料库是 M 篇文档的一个集合，表示为$D = \{W_1, W_2, \ldots, W_M\}$

LDA 目的是找到语料库上的一种概率模型，使得不仅在该语料库内的文档上取得高概率值，而且在其它相似的文档上也有很高的概率值。

LDA 的基本观点是：文档由随机混合的潜在主题表示，而每个主题是单词上的一个分布来描述的。

LDA 关于语料库 $\boldsymbol{D}$ 中每篇文档 $\boldsymbol{W}$ 的生成过程如下：

1. 选择单词数 N 服从泊松分布，$N \sim Poisson(\xi)$,
2. 选择θ服从狄利克雷分布，$\theta \sim Dir(\alpha)$.
3. 对于 N 个单词中的每个单词 w_n：
   (a) 选择一个主题 z_n，服从多项分$Mult(\theta)$.
   (b) 以概率$p(w_n|z_n,\beta)$生成单词 w_n，其中$p(w_n|z_n,\beta)$表示在主题 z_n 上的条件多项式概率.

其图模型如图 2-2 所示：

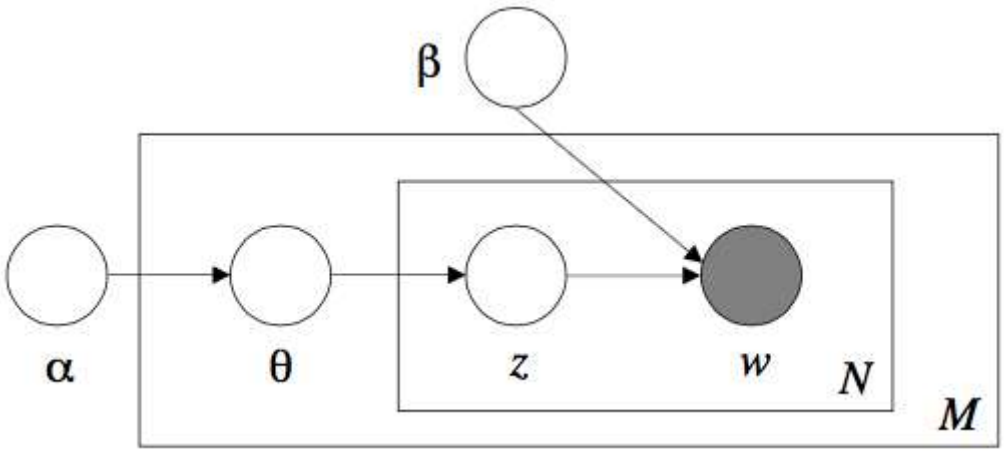

图 2-2 LDA 图模型

2.3 跨域推荐

在第一章中我们阐述了跨域推荐的概念，本节我们对于跨域推荐做一些形式化的阐述，之后在第三章提出我们自己的跨域推荐模型[46, 47]。

我们假设两个域：A 和 B，为了方便描述，下面阐述一些定义

R_A, R_B：用户评分矩阵；

U_A, U_B：用户集合；

I_A, I_B：物品集合；

$I_{AB} = I_A \cap I_B$：被两个域上的用户都评过分的物品集合；

$U_{AB} = U_A \cap U_B$：在两个域上都评过分的用户集合；

$\overline{I_A} = I_A \backslash I_{AB}$和$\overline{I_B} = I_B \backslash I_{AB}$：严格属于单一域的物品集合；

$\overline{U_A} = U_A \backslash U_{AB}$和$\overline{U_B} = U_B \backslash U_{AB}$：严格属于单一域的用户集合。

我们阐述两个主要因素：数据重叠和推荐目标。

2.3.1 跨域推荐目标

关于跨域推荐的目标，我们将其区分为三个目标：

- **单一域**：向U_A中的用户推荐I_A中的物品，域 A 将数据整合进来，例如，将域 B 的评分整合进来从而提高推荐效果。同样的也可以将域 A 的数据整合到域 B 当中。
- **跨域**：向U_A中的用户推荐I_B中的物品（同样也可以向U_B中的用户推荐I_A中的物品）。我们可以将这个目标分成两种情况：
  1. 不完全跨域：只推荐$I_{AB} \subseteq I_B$中的物品。这样它同样可以被看作是单域的情况
  2. 完全跨域：完全意义上的跨域推荐，我们向用户推荐全新的，不熟悉的物品，也就是说完全跨域在于向U_A中的用户推荐$\overline{I_B}$中的物品。
- **多域**：向$(U_A \cup U_B)$中的用户推荐$(I_A \cup I_B)$中的物品。

2.3.2 跨域数据重叠

根据两个域上用户和物品数据的重叠情况，我们将跨域场景分为了四种，如图 2-3 所示：

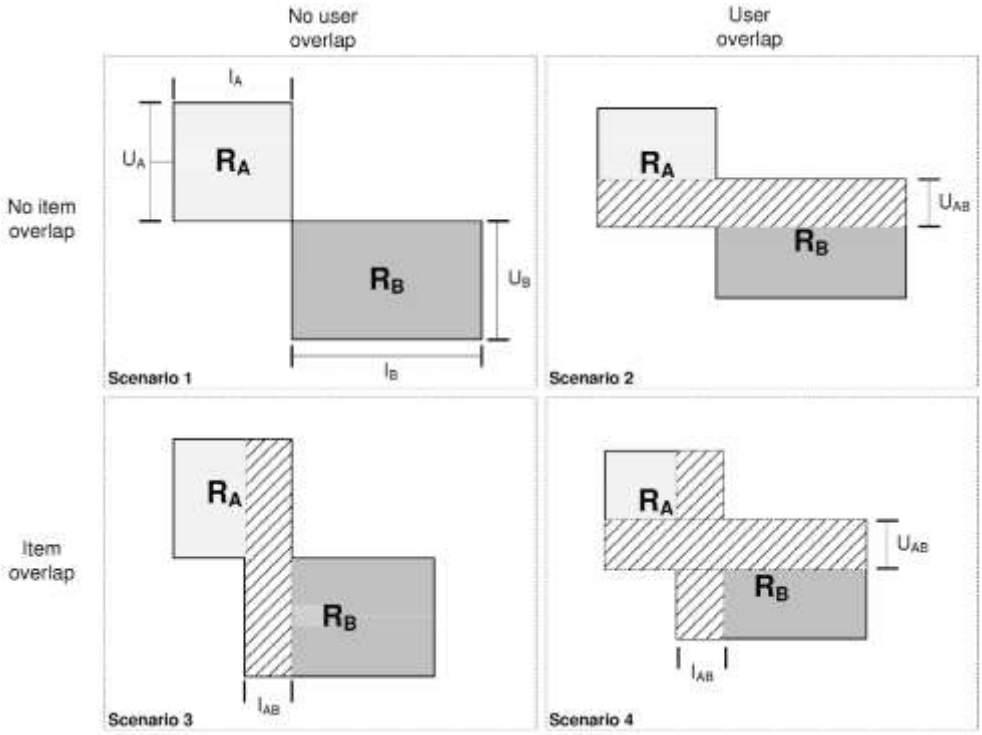

图 2-3 四种跨域用户物品数据重叠

- **无重叠**，在两个域的数据之间没有重叠部分。例如$U_{AB} = \emptyset$和$I_{AB} = \emptyset$。
- **用户重叠**，有一些用户在两个域上都有评分。例如$U_{AB} \neq \emptyset$，一些用户既喜欢看书又喜欢看电影，那么在数据集中这部分在书和电影两个域上都会有评分。
- **物品重叠**，有一些物品被两个域上的用户都评分过。例如$I_{AB} \neq \emptyset$，一台电脑可能对于某些用户来说是娱乐设施而对于另一部分用户属于办公用品，那么在电脑在两个不同的范畴当中会被不同域上的用户评分。
- **全重叠**，即用户数据和物品数据两个域上都有重叠，$I_{AB} \neq \emptyset$且$U_{AB} \neq \emptyset$。

本节我们用形式化的方式阐述了跨域推荐的场景以及推荐目标，而本文所提出的基于隐式因子和隐式主题的跨域推荐模型是在用户重叠的数据集上进行建模。

2.4 本章小结

本章基本按照推荐算法研究的时间顺序介绍了本文模型涉及到的相关技术，从最基本的近邻协同过滤算法，概率矩阵分解模型，基于主题模型的推荐模型，到当今研究的热门之一：跨域推荐的相关概念。本文的核心思想即是将跨域推荐的思想融入到经典的基于主题模型的算法当中，从而提高模型的效果以及从一定程度上对于稀疏问题和冷启动问题的解决。

第3章 基于隐式因子和隐式主题的跨域推荐算法

现存的推荐模型在对用户数据进行建模时，往往仅考虑单一域内的用户数据，而忽略了用户数据在不同域之间的联系，比如我们向用户推荐电影，那么模型仅仅考虑用户关于电影的历史行为信息，从电影这一单一域挖掘信息从而进行推荐。本文在 PMF 模型的基础上，建立潜在向量与主题因子之间的映射，并将跨域推荐的思想引入到我们的模型当中，以期达到更好的推荐效果，并且由于将跨域的思想融入到文章当中，我们的模型从一定程度上对冷启动和稀疏问题进行了解决。

3.1 问题定义

在推荐模型中，我们一般是围绕用户-物品评分矩阵进行建模的，对于单域的矩阵形式如下：

表 3-1 单一域 用户-物品评分矩阵 1

	V1	V2	V3	V4	V5
U1	3		4	2	5
U2		4			1
U3			5		
U4	3			2	

表 3-1 中给出的是评分矩阵比较常见的一种形式，用户就每一个物品给出（1，K）之间的一个评分，形成一个用户物品的评分矩阵（表 3-1 中是 5 分制）。我们在预测的时候会补全用户对于新物品的评分，然后根据评分的高低做出对于用户的推荐。

表 3-2 单一域 用户-物品评分矩阵 2

	V1	V2	V3	V4	V5
U1	1		1	1	1
U2		1			1
U3			1		
U4	1			1	

另一种比较常见的形式如表 3-2，仅仅记录用户对物品的使用情况，如果用户对物品曾经有标记、收藏等行为，就标记为 1。而在预测的时候，我们也是预测用户是否会对新物品收藏或者标记等。

以上是针对单一域的用户-物品评分矩阵，而对于跨域推荐我们需要关注用户在不同域上的行为记录，如下

表 3-3 跨域 用户-物品评分矩阵

T1	T2	T3	T4		A1	A2	A3	A4
				U1		2		
				U2	3			
	4			U3			1	
		1		U4		5		
	3			U5				4
2			5	U6				

在表 3-3 中是用户在两个域上的历史行为，我们分别定义为目标域（target）和辅助域（auxiliary），其中表中展示的是以 5 分制的用户对物品的打分。而本文就是基于类似的评分矩阵进行建模推荐的。

表 3-4 物品-单词矩阵

	W1	W2	…	W3
A1	2	1	…	0
A2	3	6	…	4
…	…	…	…	…
	W1	W2	…	W3
T1	1	0	…	5
T2	7	2	…	3
…	…	…	…	…

对于每件物品，用户在使用前和使用后都有可能去进行评论，我们可以充分利用这些评论信息。表 3-4 中展示了利用物品-单词矩阵来表现用户对物品的所有评论信息，其中每一项表示某个单词出现在某一物品评论中的次数。

基于以上阐述，本文讨论的核心内容便是如何充分利用不同域之间的用户评分信息以及用户的评论信息，从而提高在目标域上的推荐效果。

3.2 单一域建模

我们首先在单一域上利用隐式因子和隐式主题模型（Hidden Factors as Topics, HFT）[33]进行建模，建立潜在向量与主题因子之间的映射。

3.2.1 标准潜在因子模型形式

标准的潜在因子模型[48]关于用户 a 对物品 b 的评分预测是根据公式(3.1)给出

$$rec(u,i) = \alpha + \beta_u + \beta_i + U_a \cdot T_b + \lambda \|U_a\|^2 + \lambda \|T_b\|^2 \quad (3.1)$$

其中α是一个偏移量，β_u和β_i分别是用户和物品的偏移量，U_a和T_b是 K-维用户和物品潜在因子。直观的说，T_b可以看作是物品 b 的“特征”而U_a可以看作是用

户 a 对那些特征的“喜好”。给定一个训练分数集合 T，通常通过训练参数$\Theta = \{\alpha, \beta_u, \beta_i, U_a, T_b\}$，使得均方差（Mean Squared Error, MSE）最小。如：

$$\Theta = \arg\min_{\Theta} \frac{1}{|T|} \sum_{r_{a,b} \in T} (rec(u,i) - r_{a,b})^2 + \lambda \Omega(\Theta) \qquad (3.2)$$

其中$\Omega(\Theta)$通常是一个正则项，如ℓ_2正则$\|U\|_2^2, \|T\|_2^2$。有很多方法来对公式(3.2)进行优化，如最小二乘法或者梯度下降的方法。

3.2.2 添加主题模型映射

HFT 模型不像其他监督主题模型，其主题学习是与外部变量相关[49]，就像我们在第二章中关于主题模型的讨论，而 HFT 模型的发现主题学习其实是与用户和物品的隐式因子U_a, T_b有关的。

主题模型是在文档上进行操作的，所以我们首先定义我们模型中的“文档”概念。由于我们在模型中会使用用户的评论信息，所以很自然的我们可以将用户的每条评论看作是一篇“文档”$d_{a,b}$（用户 a 关于物品 b 的评论），而同样我们可以选择定义一篇文档为d_a表示用户 a 的所有评论或者d_b表示关于物品 b 的所有评论。而本文采用一个特定物品 b 的评论集d_b作为一篇文档，因为通常用户对商品做出评价的时候，考虑更多是商品本身的属性而非其个人的喜好因子。

将文档以这种方式定义，我们可以对每个物品 b 学习到一个主题分布θ_b，这个主题向量将关于物品 b 的所有评论信息在 K 个主题上的分布情况刻画了出来，其中我们假设了评分因子与评论主题的维度是相等的。

本文在对单一域进行建模的时候，利用了 HFT 中的观点：不希望评分参数T_b和评论参数θ 是独立开的，所以我们希望在建模的过程当中将两者关联起来。从直观的角度，评分因子T_b可以看作是物品 b 所拥有的属性，而用户根据自己的喜好因子U_a如果喜欢这些属性，自然会给物品 b 一个高的评分；同时在另一方面，主题θ_b定义了一个物品评论信息中的单词分布情况。所以为了将两者联系起来，

我们希望如果一个物品展现出一种特定的属性（$T_{b,k}$分量较高），那么其同时会有一个特定的主题（$\theta_{b,k}$分量较高）相对应。

然而在选取这种转换关系的过程当中，我们不能直接简单的将两者定义为相等。严格的说，θ_b是一个随机向量，其每一个分量都描述了某一主题的概率值，然而评分向量T_b可以取得R^K上的任何值。

如果我们简单强制T_b为随机向量，那么我们会损失评分模型的可解释性，而如果我们放宽对随机向量θ_b的限制，我们同样会损失对评论主题的概率解释性。所以我们需要一种允许$T_b \in R^K$，同时保证$\theta_b \in \Delta^K$的转换形式，并且它还是单调的以保证最大的T_b同时也是最大的θ_b。所以我们定义了如(3.3)的转换形式

$$\theta_{b,k} = \frac{\exp(\kappa T_{b,k})}{\sum_{k'} \exp(\kappa T_{b,k'})} \tag{3.3}$$

公式中的分母当中的指数项保证每个$\theta_{b,k}$分量都是正数，而公式中的分子确保$\sum_k \theta_{b,k} = 1$。通过这种方式，T_b可以看作是由θ_b定义的多项式上的一个自然参数。我们引入了参数κ来控制转换的峰值。当$\kappa \to \infty$的时候，θ_b仅仅会在T_b最大的索引位置接近一个值为 1 的单位向量；当$\kappa \to 0$的时候，θ_b接近一个均匀分布。直观的角度讲，较大的κ表明用户仅仅关注较为重要的主题，而较小的κ表明用户关注所有的主题。通过这种转换形式，我们在拟合参数的过程当中，仅仅拟合参数T_b。

当这些因子T_b转换到主题参数的时候，T_b必须在用户分数上精确建模(3.1)，同时其在语料库中的表现也要正常，我们将最终单一域上的模型定义如下：

$$f(T \mid \Theta, \Phi, \kappa) = \sum_{r_{u,i} \in T} (rec(u,i) - r_{u,i})^2 - \mu l(T \mid \theta, \phi) \tag{3.4}$$

其中T表示语料库，$\Theta = \{\alpha, \beta_u, \beta_i, U_a, T_b, \kappa\}$和$\Phi = \{\theta, \phi\}$分别表示分数和主题的参数，κ控制公式(3.3)的转换。公式的第一部分是公式(3.1)的预测评分误差项即在 PMF 模型的评分误差项，而公式第二部分是评论语料的似然函数。μ是一个超参数用来平衡两项之间的权重关系。

HFT 模型的图模型如图 3-1 所示：

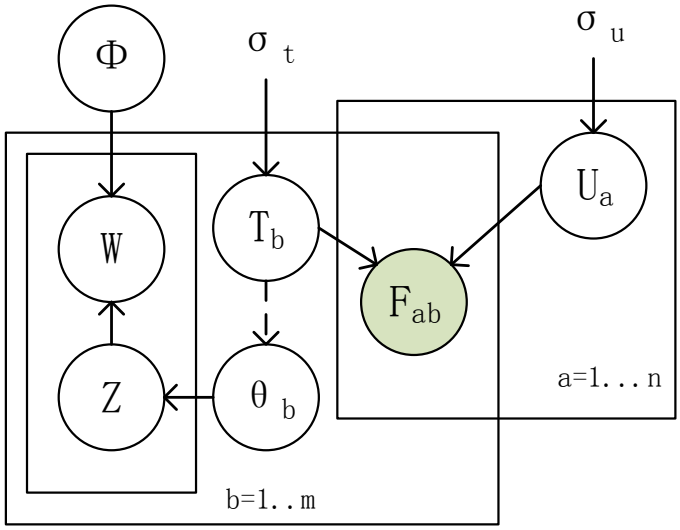

图 3-1 HFT 图模型

从图模型中我们可以看出，图中左边即是 LDA 模型，而右边则是 PMF 模型，而不同于其他如 CTR 模型，我们利用的 HFT 模型 θ_b 并不是简单的加在物品向量 T_b 上，而是采用了一种映射关系。

利用 HFT 模型的观点，我们为了更好的利用用户的反馈信息（用户对物品的评论信息），我们采用了一种将物品因子到主题向量之间的转换，从而完成了对单一域上的建模。

3.3 基于隐式因子和隐式主题的跨域推荐模型

在 3.2 我们阐述了对单一域如何进行建模，本节我们首先通过一种非线性的映射关系将不同域之间的用户潜在特征向量映射起来，之后在此基础之上提出本文的模型，基于隐式因子和隐式主题的跨域推荐模型。

3.3.1 非线性的用户向量映射

在 2.3.2 中，我们阐述了在跨域推荐中，数据重叠的四种情况，本文采用的数

据集是豆瓣（国内推荐社交平台网站）上的数据，所以我们的数据集中出现更多的是用户跨域评分的情况，所以本节我们会着重阐述在不同域的潜在用户特征向量之间建立映射关系。

对于用户 a，我们假设其目标域当中的潜在特征向量为U_a，而在辅助域中的为U'_a。我们的目标便是找到这样一种映射关系使得两者可以互相转换。这里需要解释一下为什么我们要找到一种非线性的映射关系，Xin Xin 等在对 8704 个豆瓣用户关于电影和书籍的评分进行统计后，将用户关于青春电影的评分偏移按照升序分为七个组。其中一个用户关于青春电影的评分偏移是指用户关于青春电影的平均分减去其浏览过的所有电影的平均分，以此来度量用户对于青春电影的喜好程度；而用户关于投资书籍喜好程度的度量方式类似，得出图 3-2[50]：

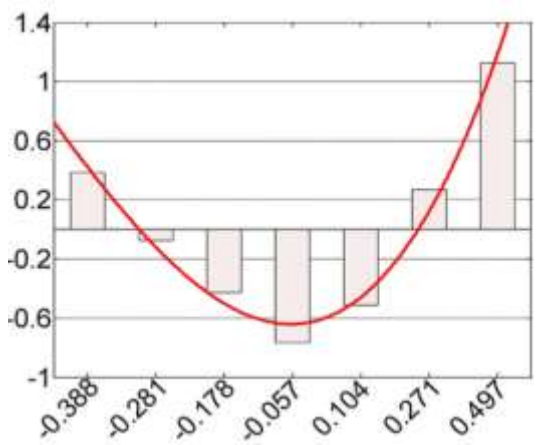

图 3-2 从青春电影到投资书籍的非线性转换形式

图中横坐标表示用户关于青春电影的平均偏移，纵坐标表示投资书籍的平均偏移。从中我们可以很明显的发现随着用户对青春电影的喜好增加，对投资书籍的喜好呈现先下降后上升的趋势，这是一种非线性的映射关系，所以我们很难找到一种可逆的从U_a到U'_a的映射，我们会试图找到两种映射函数$f(U'_a) \approx U_a$和$g(U_a) \approx U'_a$，之后我们利用这种映射关系，将一个域中的用户特征向量引入到另一个域当中。我们只讨论$f(U'_a) \approx U_a$另一个类似。

为了找到这种映射关系，我们首先假设U_a和U_a'都是 k-维向量。在实际当中，并不会要求它们是同维度的。之后将$f(U_a')$按照每个分量定义映射函数为$f^i(U_a') \approx U_a^i, i\epsilon\{1,\dots,k\}$，将$U_a'$中的第 i 维映射到U_a。我们首先定义f^i为线性的，之后再将其扩展成非线性的。

$$f^i(U_a^{'}) = (\omega^i)^T \cdot U_a^{'} + \beta^i \tag{3.5}$$

其中ω^i是向量U_a'的每一维的权重。假设 S 是用户在两个域上的反馈信息，那么$f^i(U_a')$和U_a^i之间的误差符合 0 均值的高斯分布，所以我们可以为ω^i分配一个高斯先验。之后，为了最大化映射误差的似然函数，就等价于找到参数$\{\omega^i, \beta^i\}$，来最小化带正则项的平方差。

$$\min_{\omega^i,\beta^i} \frac{1}{2}(\omega^i)^T\omega^i + \gamma\frac{1}{2}\sum_{a\in S} e_a^2, \tag{3.6}$$

$$s.t. U_a^i = f^i(U_a^{'}) + e_a, a \in S. \tag{3.7}$$

通过 KKT（Karush-Kuhn-Tucker）条件，目标函数等价于解决下面的线性问题

$$\begin{bmatrix} 0 & 1_n^T \\ 1_n & K + \dfrac{1}{\gamma} I \end{bmatrix} \begin{bmatrix} \beta^i \\ \alpha^i \end{bmatrix} = \begin{bmatrix} 0 \\ U^i \end{bmatrix}, \tag{3.8}$$

其中U^i是$|S|$维向量，其中第 a 维表示为U_a^i，而$K_{ab} = K(U_a', U_b') = \phi(U_a')^T\phi(U_b')$是 Kernel 矩阵（来自 Mercer 定理，K 为对称半正定矩阵）。通过 Kernel 技巧，K 可以被非线性函数替换，我们使用 RBF，定义如下

$$K(U_a^{'}, U_b^{'}) = \exp((\left\| U_a^{'} - U_b^{'} \right\|^2) / \sigma^2). \tag{3.9}$$

最终我们找到了非线性的f^i来表现从U_a'到U_a^i的映射，如下：

$$f^i(U_a^{'}) = \sum_{b\in S} \alpha_b^i K(U_a^{'}, U_b^{'}) + \beta^i, \tag{3.10}$$

其中ω^i被约掉，而$\{\alpha^i, \beta^i\}$是最终的参数。

3.3.2 本文模型

在我们的模型中，我们要将单一域模型(3.4)中的参数，$\{T;\ T'\}$，联合映射函数(3.10)当中的参数，$\{\alpha_m, \beta;\ \alpha'_m, \beta'\}$（为了区别式(3.4)中的全局偏移量$\alpha$，我们这里加了下标 m 予以区分），一起训练，通过固定物品在两个域上学习到的主题分布$\{\theta, \theta'\}$，我们优化联合模型的目标函数即最大化似然度，定义如下：

$$\begin{aligned} &L(\Theta,\Theta',\Phi,\Phi',\alpha_m,\beta,\alpha'_m,\beta') = \lambda L_{MAP}(U,U',\alpha_m,\beta,\alpha'_m,\beta') + \\ &(1-\lambda)(L_{HFT}(\Theta,\Phi \mid \theta,\phi) + L_{HFT}(\Theta',\Phi' \mid \theta',\phi')), \end{aligned} \tag{3.11}$$

$$L_{MAP}(U,U',\alpha_m,\beta,\alpha'_m,\beta') = \sum_{a=1}^{n} \left\| f(U'_a) - U_a \right\|^2 + \sum_{a=1}^{n} \left\| g(U_a) - U'_a \right\|^2, \tag{3.12}$$

$$L_{HFT}(\Theta,\Phi) = \sum_{r_{u,i} \in T} (rec(u,i) - r_{u,i})^2 - \mu L(T \mid \theta,\phi). \tag{3.13}$$

其中 L_{MAP} 是不同域中用户向量的映射函数，L_{HFT} 即我们在单一域上建模的 HFT 模型，λ是调整两项权重的参数。图模型如下：

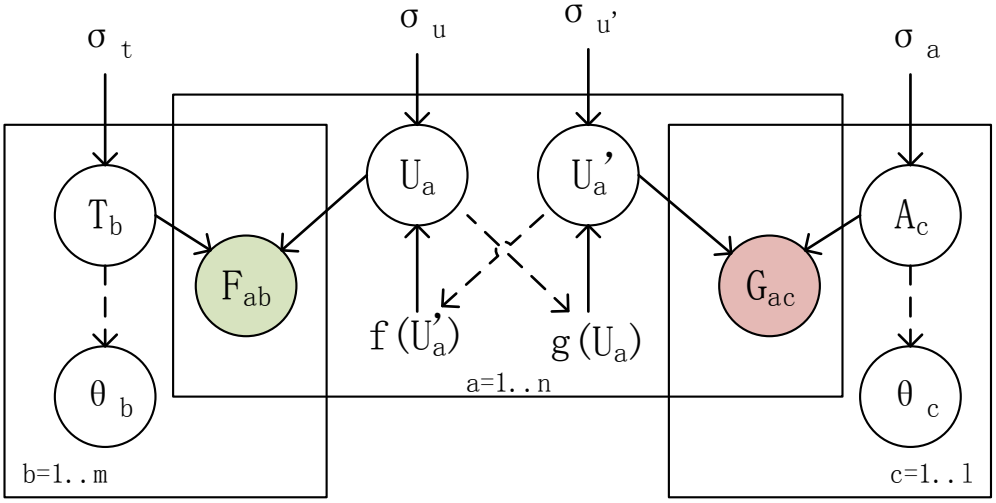

图 3-3 Cross-HFT 图模型

图模型中 T 表示目标域当中的物品潜在向量，A 表示辅助域当中的潜在物品向量，其中θ到物品向量之间我们用虚线表示，表明了这是一种映射关系而非简单

的相加。图模型是对称的，左右两边我们利用 HFT 模型分别对目标域和辅助域建模，中间 f 和 g 便是我们定义的非线性的用户向量映射函数。

3.4 模型训练

我们在模型训练时需要找到参数$\{\Theta,\Phi,\Theta',\Phi';\alpha_m,\beta,\alpha'_m,\beta'\}$来最优化目标函数(3.11)，在给定$\{\alpha_m,\beta;\ \alpha'_m,\beta'\}$，我们利用梯度下降的方法来搜索$\{\Theta,\Phi,\Theta',\Phi'\}$。其中每次迭代公式如下：

$$\begin{aligned}\nabla_{U_a}L&=(1-\lambda)\left[\sum_{b\in B(a)}(r_{ab}-rec(a,b))\cdot T_b+\lambda_a U_a\right]+\lambda(U_a-f(U'_a))+\frac{1}{\sigma_u^2}U_a\\ \nabla_{T_{bk}}L&=(1-\lambda)\left[\sum_{a\in A(b)}(r_{ab}-rec(a,b))\cdot U_a+\lambda_b T_b-\mu\kappa(n_{bk}-n_b\frac{e^{\kappa T_{bk}}}{\sum_{k'}e^{\kappa T_{bk'}}})\right]\end{aligned}\tag{3.14}$$

其中n_{bk}表示物品 b 的评论文档中每个主题出现的次数，n_b表示每个物品评论文档中的单词数，基于此，我们模型训练的基本步骤如下：

算法 1 参数估计

输入：跨域矩阵反馈信息，以及模型的超参数

输出：$\{\Theta,\Phi,\Theta',\Phi';\alpha_m,\beta,\alpha'_m,\beta'\}$

1. 初始化$\{\Theta,\Phi,\Theta',\Phi';\alpha_m,\beta,\alpha'_m,\beta'\}$
2. for 每次迭代
3. 根据梯度下降公式(3.14)更新参数$\{\Theta,\Phi,\Theta',\Phi'\}$
4. 通过解式(3.8)更新$\{\alpha_m,\beta,\alpha'_m,\beta'\}$
5. end for

其中在每次迭代的过程当中，梯度下降部分我们利用 L-BFGS 来进行计算，

它是一种用来解决非线性优化问题的牛顿方法，利用梯度下降来计算参数；而在更新主题参数的部分我们利用吉布斯采样的方式进行计算，之后我们将本次迭代计算得到的U_a, U_a'代入到式(3.8)中，利用 KKT 条件，调用 gsl 库中关于矩阵运算的部分更新$\{\alpha_m, \beta, \alpha_m', \beta'\}$。

3.5 本章小结

本章我们首先在单一域上通过在潜在向量与主题因子之间建立映射关系，利用 HFT 进行建模，之后我们通过非线性映射函数，将不同域之间的用户潜在向量关联起来，从而将不同域上的数据关联起来，提出了我们自己的基于隐式因子和隐式主题的跨域推荐模型。最后我们用图模型的方式阐述了整个模型，并且就模型训练的过程做了基本的阐述。

第4章 实验与分析

4.1 数据集及预处理

在实验部分我们使用的数据集来源于国内知名的社交推荐网站豆瓣网，豆瓣（douban.com）网站以书影音为主，提供关于书籍、电影、音乐等作品的信息，而无论关于作品的描述还是评论信息都由用户提供。当用户在网站注册登录后，便可以进行评论评分，写出自己的读后感、观后感，并给出 5 分制（1-5 整数）的一个分数。我们的数据集来源于豆瓣网，是由 zhong[51]爬取和提供的，包括：书籍 144276 本，电影 39962 部，音乐 86429 首，用户 31514。

由于本文的跨域模型是在两个域上进行建模，我们在调整参数阶段将实验设定在书籍和电影两个域上进行，之后我们也会在书籍和音乐以及电影和音乐两个跨域上进行实验，比较不同跨域对实验效果的影响，最终我们在进行对比实验时，会在三个域上进行不同模型间的对比。三个域共有 300865 条评论（包括评分），其中书籍上的评论有 11492 条，电影上的评论有 15417 条，音乐域上的评论有 872 条。为了方便描述我们将有历史行为信息的用户称为“活跃用户”，我们对数据集做了统计如下：

表 4-1 数据集统计表

类别	数量
总用户数	31514
书总数	144276
电影总数	39962
音乐总数	86429
书评论总数	11492
电影评论总数	15417
音乐评论总数	3176
书籍域活跃用户数	2364
电影域活跃用户数	2283
音乐域活跃用户数	872
跨域活跃用户数（书籍/电影）	1338
跨域活跃用户数（书籍/音乐）	553
跨域活跃用户数（电影/音乐）	631

其中跨域活跃用户是指在不同域间都有历史行为信息的用户。为了方便描述，我们用 B 表示书籍域、M 表示电影域、MU 表示音乐域。如图 4-1 中展示的三个域之间活跃用户的关系

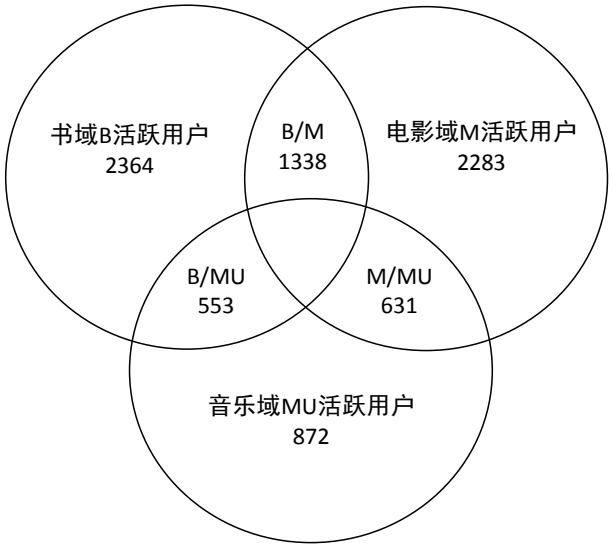

图 4-1 活跃用户关系图

在三个域的数据集上，我们都以 HFT 模型为基础建立模型，而对于跨域活跃用户来说，我们在这些用户上建立映射关系作为L_{Map}添加到模型中，如公式(3.11)中的那样。

对于数据的预处理，由于我们采用的是豆瓣数据集，其中用户的评论部分为中文，我们首先对评论做了分词，并且将评论当中的停止词去掉，再将其作为输入放入到我们的模型中进行建模。在数据预处理阶段，因为只需对每条数据记录进行遍历，时间复杂度为 O(n)。

4.2 实验设计

我们在进行实验设计的过程中，主要从两个角度入手，一方面是模型本身的角度，即我们对模型参数的调试过程；另一方面是模型与其他先进的模型比较，对比模型之间的优劣。所以我们的实验设计分为以下三个步骤：

- 在书籍和电影域上对模型的参数进行训练与调整；
- 对比书籍和电影域、书籍和音乐域、电影和音乐域三个跨域中，模型的表现；

- 选取其他模型在三个域上分别与我们模型的实验效果进行对比。

4.2.1 数据集划分

我们的数据集划分为两个域，我们分别将两个域上的数据集按照比例 80%，10%和 10%进行划分，之后在两个域的数据集上进行建模和训练，最后各自独立的进行最终结果的评价。

4.2.2 评价标准

常用的评价标准有很多，如准确率（precision）、召回率（recall）、F-Score、平均绝对误差 MAE（Mean Absolute Error）、均方差 MSE（Mean Square Error）等。

准确率和召回率[52]来源于信息检索领域，用来描述信息检索的效果，而在推荐领域也可以用它来描述对用户推荐效果的判断，其计算方式如下：

$$precision = \frac{T_P}{T_P + F_p}$$
$$recall = \frac{T_p}{T_p + F_N} \tag{3.15}$$

其中准确率表示表示正确推荐给用户的物品数T_p（即推荐给用户并且是用户喜欢的物品数）占所有实际推荐的物品总数$T_p + F_p$的比例。而召回率表示正确推荐给用户的物品数T_p占所有应该被推荐给用户的物品总数$T_p + F_N$的比例。准确率和召回率都是我们模型推荐给用户的物品与用户真实喜好程度的一个比值，所以准确率和召回率都是越高，模型推荐的效果越好。

而 MSE 通常用来度量通过模型预测值或估计值与真实观测值的误差，在推荐系统中，当我们通过模型对用户的评分做出预测后，经常使用 MSE 来度量我们预测分数与已知用户真实分数的误差，其运算公式如下：

$$MSE = \frac{\sum_{ij}(r_{ij} - rec(i,j))^2}{N} \tag{3.16}$$

类似的评价标准 MAE[53]也是从预测分数与实际用户打分的误差来评价模型的优劣程度，其计算方式如下：

$$MAE = \frac{\sum_{ij}\left|r_{ij} - rec(i,j)\right|}{N} \tag{3.17}$$

其中r_{ij}表示测试集中用户 i 对物品 j 的真实评分，而$rec(i,j)$即我们通过推荐模型预测出的用户评分。MSE 会对误差求平方，而 MAE 是对误差求绝对值，两者比较类似。MSE 和 MAE 值越小，说明模型预测出的分数与用户真实的评分数据越接近，那么模型越优秀，而反之则越差。

由于本文采用的豆瓣数据集采用的是 5 分制（1-5 整数）的评分方式，此外 MSE 评价方式在防止过拟合时比 MAE 更加灵敏，所以我们在进行模型参数训练时采用 MSE 来衡量模型推荐效果。另外，在对比实验中，我们为了更加全面的衡量各个模型，也比较了模型之间 MAE 的大小。

4.2.3 参数训练

由于模型中涉及到两个域的参数，我们将另一个域的参数上面加'。模型中涉及到的超参数如表 4-2 所示：

表 4-2 超参数列表

参数	描述
λ, λ'	式(3.1)用户和物品潜在向量正则项参数
μ, μ'	式(3.4)中主题模型部分的权重参数
K	用户和物品潜在向量的维度
λ_{global}	式(3.11)中跨域映射和 HFT 的权重关系
σ^2	式(3.9)映射函数的先验参数
γ	式(3.8)KKT 条件中防止矩阵 K 奇异的参数

其中参数γ, σ^2根据文章[50]中所提及的，我们作为经验参数，设为 500 和 2.5，另外由于我们在两个域上都是利用 HFT 模型进行建模，所以参数λ, λ'和μ, μ'我们以相同的值来进行调参（当然，这里仅仅是调整参数的过程，最后进行实验的时候可以取不一样的值），我们的参数训练分成三个步骤进行：

- 首先我们通过改变 K 的值来观察模型的效果，首先确定 K 的值；
- 固定参数 K 和参数λ_{global}，对两对参数λ, λ'和μ, μ'进行调参；
- 最后在参数λ, λ'和μ, μ'调好后，我们固定 K 对参数λ_{global}进行调参。

4.3 实验结果与分析

本节我们按照实验设计中参数训练的步骤，分别对每组参数的对模型的影响进行分析，最终将模型调整到最优状态。

4.3.1 K 值的影响分析

K 作为潜在向量维度的参数，对于模型的推荐效果和运算效率都会有影响，当我们的 K 值较小时，模型运算过程会比较快，然后可能在推荐效果方面会有所下降，所以我们需要在两者之间做一个折中。我们根据经验首先将λ, λ'和μ, μ'设置为 35 和 5，将λ_{global}设置为 0.1，对 K 进行实验。

表 4-3 Cross-HFT 不同 K 值的 MSE

Domain/K	5	10	15	20
Movie	0.890675	0.883877	0.876624	0.874712
Book	0.990681	0.986762	0.978871	0.977284

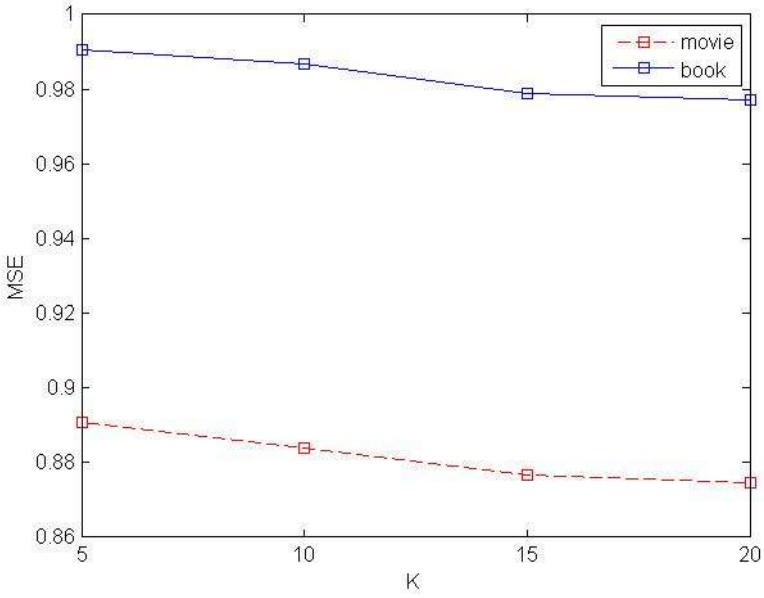

图 4-2 K 值 MSE 变化曲线

从图中我们可以看出随着 K 值的增长，两个域的推荐效果呈现不同的趋势：在 movie 域上，随着 K 值的增大，整体 MSE 呈现下降趋势，即推荐效果在变好，不过在 K 继续增大，MSE 下降趋势也趋于平缓；另一方面，在 book 域上，随着 K 值的增大，MSE 呈现了一定的波动，不过随着继续增大，MSE 也呈现下降的趋势。而在本实验中，随着 K 值的增大，整个模型的运行效率会有所下降，为了在模型效果和运行效率上进行折中，将 K 的值设置为 15。

4.3.2 参数λ, λ'和μ, μ'对模型影响分析

在 K 固定为 15 后，我们开展下面的实验。调整λ, λ'和μ, μ'两对参数，观察推荐效果的变化。下面我们针对用户在电影和书上的数据集分别展开实验。

表 4-4 域 1 不同λ,μ值的 MSE

$\lambda\backslash\mu$	1	5	10	15	50
1	0.926898	0.902414	0.886809	0.897413	0.918426
5	0.895366	0.878218	0.877488	0.886268	0.926084
10	0.891569	0.872783	0.876305	0.884639	0.92708
15	0.881613	0.876396	0.874572	0.873605	0.924225
50	0.881889	0.881673	0.878547	0.882416	0.938202

表 4-5 域 2 不同λ',μ'值的 MSE

$\lambda'\backslash\mu'$	1	5	10	15	100
1	1.125203	1.082701	1.038842	1.058626	1.059751
5	1.019984	1.108227	1.041623	1.049815	1.102966
10	1.03277	1.093084	1.006366	1.035324	1.075661
15	0.998723	0.998436	1.010218	1.025611	1.094943
100	0.985139	0.985339	1.013615	1.015762	1.081978

我们在跨域推荐的过程中，分别记录了电影域和书籍域上的 MSE 情况，为方便我们找到最优参数的位置，我们以等高线图刻画了两个参数变化对模型结果的影响，如图 4-3 与 4-4 所示：

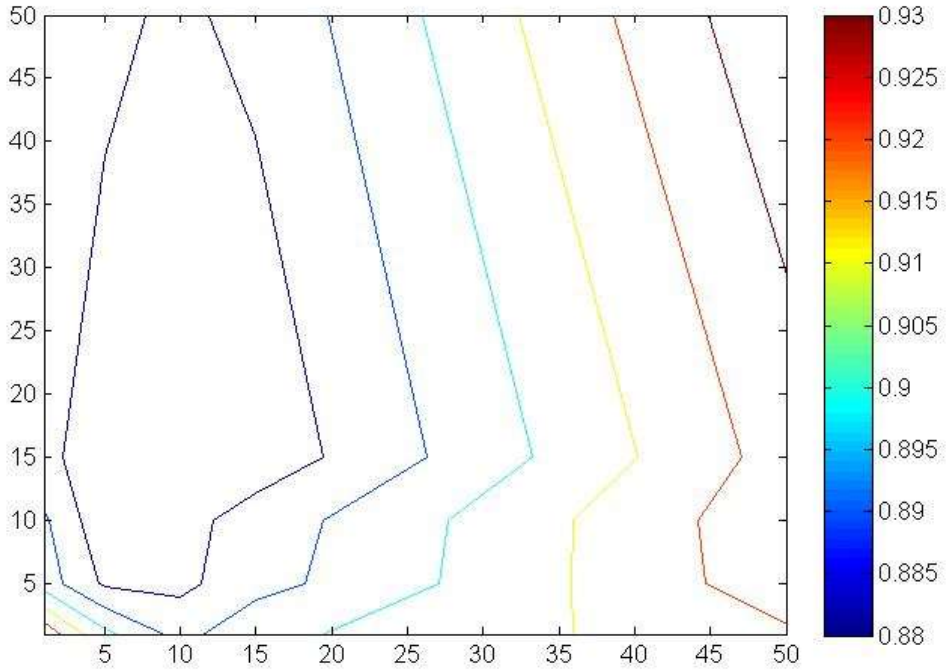

图 4-3 $\lambda\backslash\mu$ 值对 MSE 影响的变化等高线

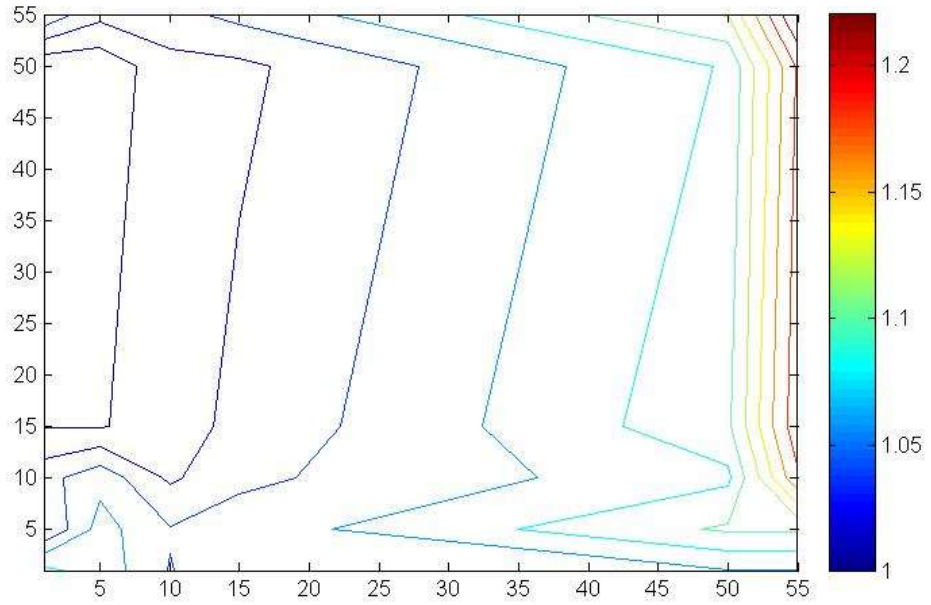

图 4-4 $\lambda\backslash\mu'$ 值对 MSE 影响的变化等高线

由于我们采用 MSE 作为衡量推荐效果的方式，所以当 MSE 越低时，模型效果越好，从图中我们可以看出在深色部分区域，MSE 值最小，则模型整体推荐效果最优秀，根据等高线图给出的参数范围，通过多次实验，我们可以得出当λ, λ'和μ, μ'分别取在 50 和 3 附近时模型取得最好的效果。

4.3.3 参数λ_{global}对模型影响分析

在调整了上面的参数后，我们调整参数λ_{global}来查看对推荐效果的影响，λ_{global}作为调整L_{Map}项和L_{HFT}项之间的权重，我们可以控制跨域映射对于整个模型的影响。

表 4-6 不同λ_{global}值的 MSE

Dataset\λ_{global}	0.1	0.3	0.5
Movie	0.873390	0.878810	0.895409
Book	0.979217	1.012023	1.013425

我们为了更加形象的观察λ_{global}对实验效果的影响，我们将实验结果数据以折线图的形式呈现出来。如图 4-5 所示：

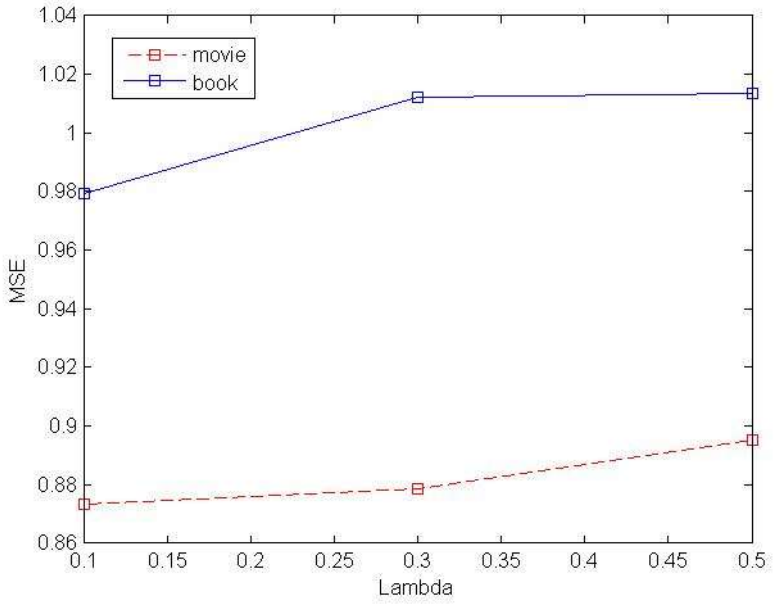

图 4-5 λ_{global}值 MSE 变化曲线

从图中我们可以观察出随着λ_{global}增大，MSE 的值也在增大，即整个模型的推荐效果有所下降。而从模型的角度考虑，由于λ_{global}是控制映射项与 HFT 项之间的权重关系，如公式(3.11)，所以通过实验，我们可以理解为，映射项在反馈信息中所占的比重不能过大，这样整个模型的推荐效果较好。所以我们将λ_{global}取在 0.1 附近。

4.3.4 Top 词分析

在最开始进行实验的时候，我们发现每个主题的最高的 10 个词汇是一些停止词，如“的”，“这”，“一些”等等，所以我们在数据预处理的时候，专门对停止词进行了过滤，之后我们在 *K=5* 的时候，在电影、书籍和音乐域的数据集上得出了下面的 Top 词表。如表 4-7、4-8 和 4-9 所示：

表 4-7 电影每个主题的 Top10 词汇（*K=5*）

Topic1	Topic2	Topic3	Topic4	Topic5
戒	教授	姜文	赤壁	长江
李安	机器人	升起	吴宇森	延续
色	萨	获奖	三国	单身
值得一看	谎言	孝	經	七号
七剑	雅	照常	洋	山田
无极	天才	外星人	反派	南京
名状	樟	电视剧	一口	想看
佳	爱丽丝	太阳	記	星
芝	原则	围	那一刻	映
张爱玲	未来	dv	黑社会	伊朗

表 4-8 书籍每个主题的 Top10 词汇（*K=5*）

Topic1	Topic2	Topic3	Topic4	Topic5
天空	书	红	好啦	奖
疯狂	不像	雅	究竟	传播
规则	想想	爸爸	哲学家	童
有种	却又	非	城	夏
浪漫	热	一路	高中	实践
悲伤	费	错误	間	经典
科幻小说	看似	原文	案例	大师
恐怕	宋	最喜欢	文明	整理
就会	民主	出于	据说	原著
幸运	入门	索	比喻	例子

表 4-9 音乐每个主题的 Top10 词汇（*K=5*）

Topic1	Topic2	Topic3	Topic4	Topic5
声音	歌手	爱	这张	喜欢
the	流行	感觉	歌	里
首歌	许巍	一直	碟	说
好	听过	这种	cd	已经
歌	唱	摇滚	风格	这是
听	作品	喜欢	觉得	现在
感觉	想	最	的人	听
人	一个	专辑	说	听到
一首	出了	张	后	东西
听了	张	http	音乐	名字

在表中我们发现的主题是容易解释和理解的。发现主题与对物品划分类别种类是相似的。而从模型的观点，一个简单的说明来解释为什么我们的模型发现不同种类的主题：首先，用户更倾向于对相似种类的物品进行评分，所以将物品划分为不同的种类解释了在评分数据上的不同。其次，不同的语言、词汇被用来描述不同种类的物品，所以种类同样解释了用户对于物品的评论数据的不同。

4.3.5 不同跨域对模型效果的影响分析

由于我们的数据集分为书籍、电影和音乐三个域，而我们的跨域模型是基于两个域进行建模的，所以接下来我们以两个为一组，形成三个跨域，分别利用我们的模型进行实验，并分析实验结果。如表 4-10 所示：

表 4-10 不同跨域 MSE

Res\Cross	M\B	B\MU	M\MU
MSE	0. 872783\0.979217	1.033203 \ 0.814304	0.874864\ 0.829672

为了我们比对数据，我们将三组数据分三个类别放在柱状图中，如图 4-6 所示：

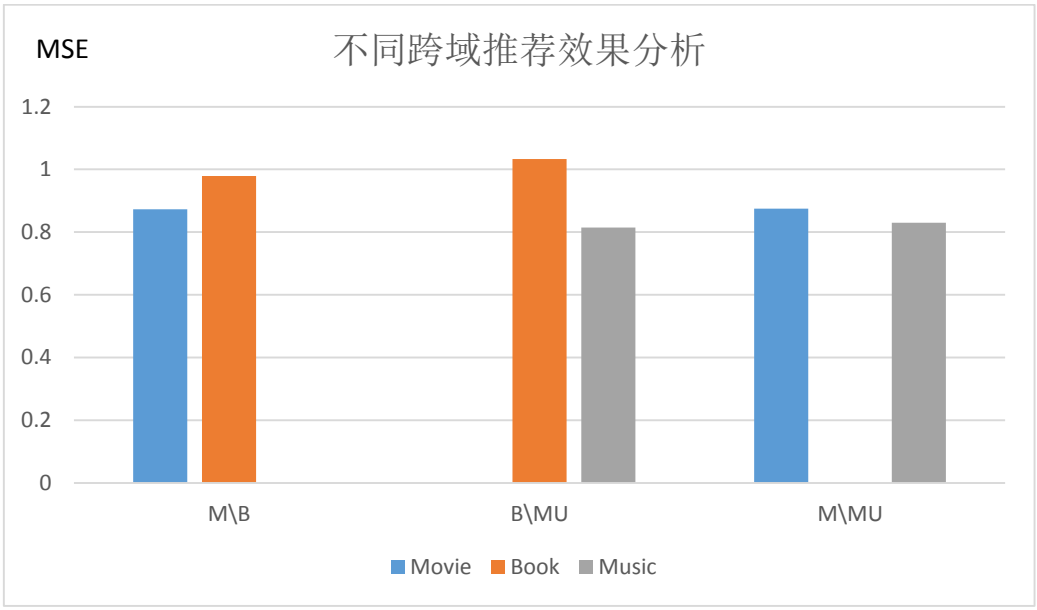

图 4-6 不同跨域 MSE

从图中我们可以看出，利用不同跨域组合来进行推荐效果是不同的，我们发现电影和书籍域在一起进行跨域推荐时达到效果最好，而与音乐域进行联合跨域推荐时，效果差一些；而对于音乐域，当其与书籍域进行联合推荐时要好于与电影域进行联合跨域推荐。这一方面与我们的数据集有关，另一方面也与域本身的性质有关。

4.4 对比实验

为了验证本文模型的推荐效果，我们找了目前优秀的推荐模型来进行比较。首先我们与本文涉及到的经典模型进行比较：CF 模型，即传统的近邻协同过滤模型；PMF 模型，概率矩阵分解模型； HFT 模型，在主题模型因子与潜在向量之间建立映射关系，来将用户评论信息引入到 PMF 模型中。另外，我们与 15 年 XinXin 等人提出的 Cross-Domain Collaborative Filtering with Review Text（Cross-CTR）模型进行了对比，我们从 MSE 和 MAE 两个角度分别在电影、书籍和音乐

三个域上来衡量各模型之间的优劣程度，如表 4-11 和 4-12 所示：

表 4-11 模型间 MSE 的比较结果

dataset	CF	PMF	HFT	Cross-CTR	Cross-HFT
Movie	1.246782	1.024391	0.882918	0.880831	0. 867526
Book	1.48315	1.253728	1.035367	1.025739	0.977017
Music	1.154893	0.987425	0.849715	0.830579	0.814304

表 4-12 模型间 MAE 的比较结果

dataset	CF	PMF	HFT	Cross-CTR	Cross-HFT
Movie	1.036383	0.954135	0.712880	0.712361	0.705836
Book	1.097641	0.984739	0.750065	0.738025	0.720701
Music	1.057458	0.896014	0.725194	0.710683	0.701051

为了更好的进行对比，我们采用柱状图的形式将实验结果进行了统计与对比，并且由于我们采用 MSE 和 MAE 作为评价标准，实验结果越低则模型效果越好，为了更加形象，我们对实验结果求倒数再放入柱状图中进行比较，则柱状图越高表明实验结果越好。如图 4-6 和图 4-7 所示：

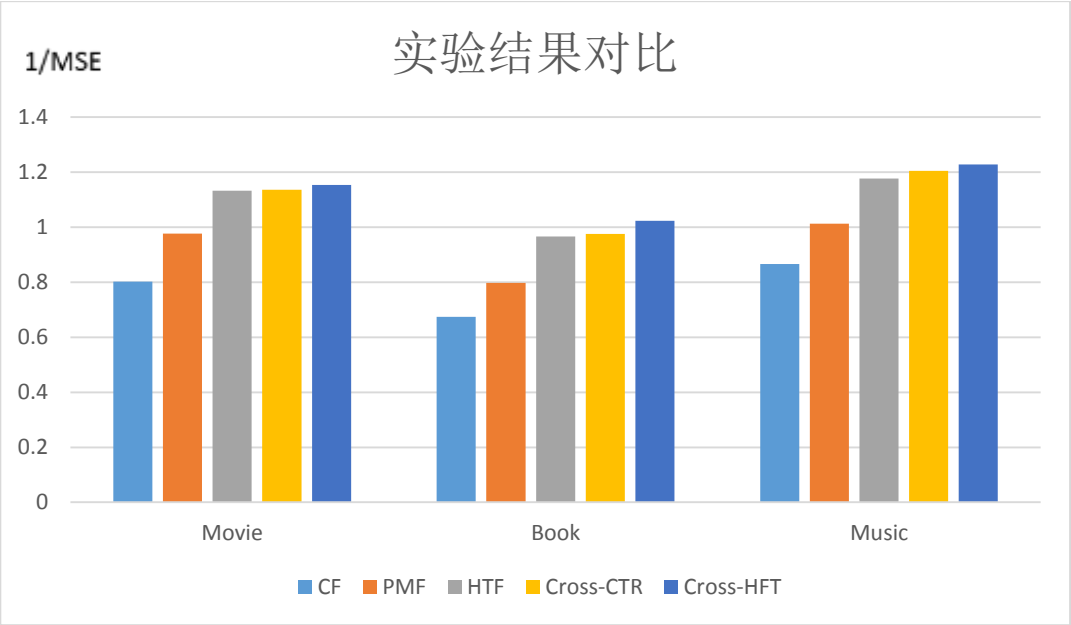

图 4-7 对比实验 MSE 结果图

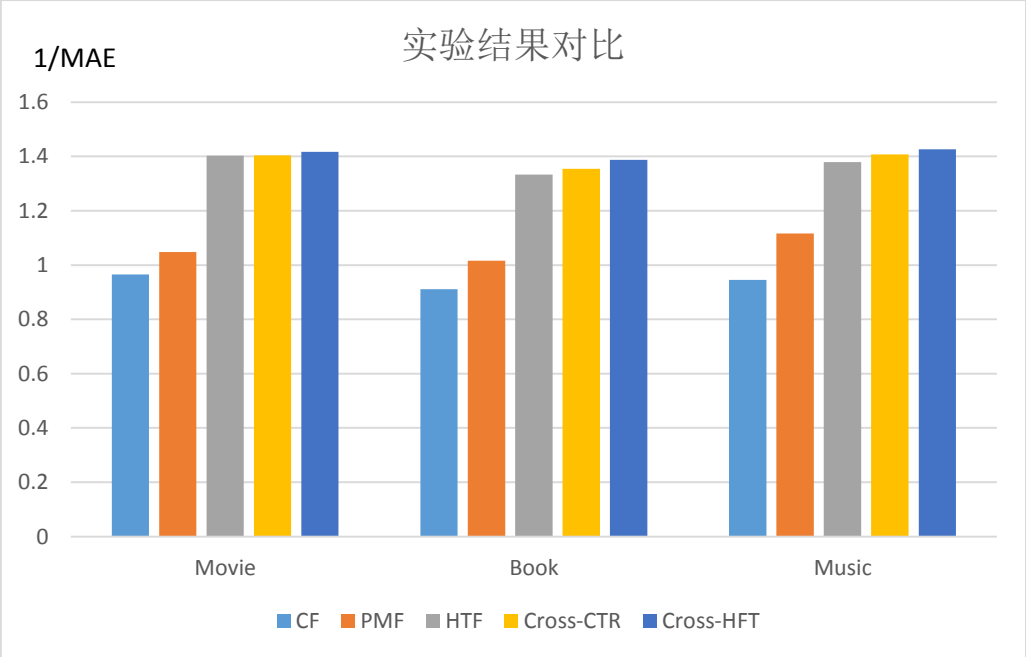

图 4-8 对比实验 MAE 结果图

从图中我们可以看出，传统的 CF 推荐模型，在比较稀疏的数据集上表现非

常糟糕；而概率矩阵分解模型，将高维度的矩阵空间分解为低维度的潜在向量，在表现上明显优于 CF 模型。在此基础上，将评论信息以主题模型进行建模，作为反馈信息加入到传统的 PMF 模型当中，使得模型的效果有了显著的提高；而对于最后两个跨域类型的推荐模型，由于都将跨域的思想引入到了推荐模型当中，可以发现，其推荐效果要优于前面几种模型。而对于这两种模型而言，Cross-CTR 模型在利用用户的评论信息时，是直接利用主题模型进行建模，之后将主题因子直接加入到 PMF 的潜在向量中；而不同的是，我们提出的 Cross-HFT 模型，在利用主题因子的时候，是建立了一种评分参数T_b和评论参数θ之间的映射，如式(3.3)中的那样，所以在最终的实验结果上我们也可以看到，我们的模型优于 Cross-CTR 模型。

4.5 本章小结

本章我们首先介绍了实验数据的来源，并对实验数据进行了统计与分析。之后我们详细阐述了实验的设计方案，首先通过对不同参数的训练调整，使得模型达到最好的运行效果；之后设定了对比实验，通过与不同模型的对比与分析，我们得出了本文提出的 Cross-HFT 模型相较于以往的模型有了明显提升。

第5章 总结与展望

5.1 本文工作

推荐算法是当前研究的热点之一，其应用领域也非常广阔，平时生活中经常可见到推荐的身影。本文梳理了国内外推荐算法的研究现状，以及当前研究存在的问题。一方面，我们在经典的概率矩阵分解模型的基础上，加入了主题模型，充分利用用户的评论信息作为反馈加入到矩阵分解后的潜在向量当中；另一方面，我们在建立模型的时候，不仅仅局限在单一域上，将单一域扩展为跨域推荐，其中具有挑战的便是在不同域的潜在向量之间建立映射关系，从而达到跨域的推荐效果。实验中，我们利用梯度下降、吉布斯采样以及 KKT 条件，对模型参数进行训练求解，并通过参数调试记录模型的不同推荐效果，而对比实验也表明，我们的模型在推荐效果上有了一定提高。

本文工作总结如下：

（1） 介绍了推荐系统的研究背景与研究意义，梳理了国内外关于推荐领域的研究现状以及存在的问题与缺陷。

（2） 详细阐述了本文涉及到的相关技术点，包括最基本的近邻协同过滤，经典的概率矩阵分解模型，之后我们将关注点转移到本文的两个核心点，即主题模型和跨域推荐相关技术。主题模型我们深入讨论了本文使用的 LDA 主题模型，而在另一方面我们总结了跨域推荐系统的相关技术。

（3） 将主题模型、跨域推荐结合到传统的概率矩阵分解模型，我们首先在单一域上用隐式因子和隐式主题模型进行建模，之后通过非线性的映射函数，将跨域引入到模型中，最终提出了我们的基于隐式因子和隐式主题的跨域推荐模型。

（4） 在对模型的训练过程中，我们首先利用梯度下降，对目标函数求偏导，之后在主题模型部分我们用吉布斯采样的方式进行求解，最后对于跨域映射函数部分，我们利用 KKT 条件的矩阵方程进行求解，经过 EM 迭代过程使得模型达到最优状态。最终在国内知名网站豆瓣的数据集上，我们进行了参数调整和对比实验。

5.2 未来工作展望

本文将跨域（Cross-domain）与主题模型结合到传统的 PMF 模型中，充分利用了用户评论信息的同时还将用户在不同域之间的信息关联起来，从而达到了更好的模型效果。然而，本文模型在提升效果的同时，在运算效率上有所下降，另外在不同域之间的映射上也可以有所变化。因此未来工作可以对以下三个方面进行探索：

（1） 本文提出的跨域模型是针对于两个域的数据进行建模，是否可以将其扩展为多域，这样在利用了每个域的数据同时也充分利用了不同域之间的关联性，可以试图去聚合不同域之间的数据，但是多域间数据映射关系的建立是一个难点。

（2） 在不同域之间进行映射的时候，我们主要针对不同域之间的用户潜在向量进行非线性的映射，是否可以尝试在物品潜在向量上进行映射，不过这需要找到合适的数据集。另一方面，我们在做映射的时候，更多的是关注到活跃用户（即在两个域都有行为信息的用户）间的映射，而对于单一域的活跃用户（即仅有单一域内的行为信息的用户）也可以尝试将其映射到另一个域上。

（3） 由于在原有 PMF 模型的基础上我们加入了很多新的反馈信息，那么在运算过程中可以尝试进行一些优化，以提高整个模型的推荐效率。

参考文献

[1] Adomavicius, G., A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(6): 734-749

[2] Linden, G., B. Smith, J. York. Amazon. com recommendations: Item-to-item collaborative filtering[J]. Internet Computing, IEEE, 2003, 7(1): 76-80

[3] Balabanović, M., Y. Shoham. Fab: content-based, collaborative recommendation[J]. Communications of the ACM, 1997, 40(3): 66-72

[4] Su, X., T.M. Khoshgoftaar. A survey of collaborative filtering techniques[J]. Advances in artificial intelligence, 2009: 1-19

[5] Bobadilla, J., F. Ortega, A. Hernando, et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46: 109-132

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

[7] Paterek, A. Improving regularized singular value decomposition for collaborative filtering. Proceedings of KDD cup and workshop, 2007: 5-8

[8] Mnih, A., R. Salakhutdinov. Probabilistic matrix factorization. Advances in neural information processing systems, 2007: 1257-1264

[9] Hu, M., B. Liu. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004: 168-177

[10]Hu, M., B. Liu. Mining opinion features in customer reviews. AAAI, 2004: 755-760

[11]Hu, M., B. Liu. Opinion extraction and summarization on the web. AAAI, 2006: 1621-1624

[12]Liu, B., M. Hu, J. Cheng. Opinion observer: analyzing and comparing opinions on the web. Proceedings of the 14th international conference on World Wide Web, 2005: 342-351

[13]Agrawal, R., R. Srikant. Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases, VLDB, 1994: 487-499

[14]Wang, K., L. Tang, J. Han, et al. Top down fp-growth for association rule mining, Springer. 2002

[15]Singhal, A. Modern information retrieval: A brief overview[J]. IEEE Data Eng. Bull., 2001, 24(4): 35-43

[16]Goldberg, D., D. Nichols, B.M. Oki, et al. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70

[17]Paterek, A. Improving regularized singular value decomposition for collaborative filtering. Proceedings of KDD cup and workshop, 2007: 5-8

[18]Yang, X., Y. Guo, Y. Liu, et al. A survey of collaborative filtering based social recommender systems[J]. Computer Communications, 2014, 41: 1-10

[19]Massa, P., P. Avesani. Trust-aware collaborative filtering for recommender systems, in On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE, Springer. 2004, 492-508

[20]Massa, P., P. Avesani. Trust-aware recommender systems. Proceedings of the 2007 ACM conference on Recommender systems, 2007: 17-24

[21]Ma, H., I. King, M.R. Lyu. Learning to recommend with social trust ensemble. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009: 203-210

[22]Jamali, M., M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. Proceedings of the fourth ACM conference on Recommender systems, 2010: 135-142

[23]Chen, C., J. Zeng, X. Zheng, et al. Recommender System Based on Social Trust Relationships. e-Business Engineering (ICEBE), 2013 IEEE 10th International Conference on, 2013: 32-37

[24]Yang, X., H. Steck, Y. Liu. Circle-based recommendation in online social networks. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012: 1267-1275

[25]Tang, J., H. Gao, H. Liu. mTrust: discerning multi-faceted trust in a connected world. Proceedings of the fifth ACM international conference on Web search and data mining, 2012: 93-102

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

[27]Chen, C., X. Zheng, Y. Wang, et al. Context-aware collaborative topic regression with social matrix factorization for recommender systems. Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014: 9-15

[28]Blei, D.M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84

[29]Blei, D.M., A.Y. Ng, M.I. Jordan. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022

[30]Wang, C., D.M. Blei. Collaborative topic modeling for recommending scientific articles. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011: 448-456

[31]Purushotham, S., Y. Liu, C.J. Kuo. Collaborative topic regression with social matrix factorization for recommendation systems[J]. arXiv preprint arXiv:1206.4684, 2012

[32]Kang, J., K. Lerman. LA-CTR: A limited attention collaborative topic regression for social media[J]. arXiv preprint arXiv:1311.1247, 2013

[33]McAuley, J., J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. Proceedings of the 7th ACM conference on Recommender systems, 2013: 165-172

[34]Loizou, A. How to recommend music to film buffs: enabling the provision of recommendations from multiple domains, Ph.D. dissertation, University of Southampton. 2009

[35]González, G., B. López, J.L. de la Rosa. A multi-agent smart user model for cross-domain recommender systems[J]. Proceedings of Beyond Personalization, 2005

[36]Tuffield, M.M., A. Loizou, D. Dupplaw. The semantic logger: Supporting service

building from personal context. Proceedings of the 3rd ACM workshop on Continuous archival and retrival of personal experences, 2006: 55-64

[37]Lee, C., Y. Kim, P. Rhee. Web personalization expert with combining collaborative filtering and association rule mining technique[J]. Expert Systems with Applications, 2001, 21(3): 131-137

[38]Berkovsky, S., T. Kuflik, F. Ricci. Cross-domain mediation in collaborative filtering, in User Modeling 2007, Springer. 2007, 355-359

[39]Zhuang, F., P. Luo, H. Xiong, et al. Cross-domain learning from multiple sources: a consensus regularization perspective[J]. Knowledge and Data Engineering, IEEE Transactions on, 2010, 22(12): 1664-1678

[40]Cao, B., N.N. Liu, Q. Yang. Transfer learning for collective link prediction in multiple heterogenous domains. Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010: 159-166

[41]Salton, G., A. Wong, C. Yang. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620

[42]Wong, S.M., W. Ziarko, P.C. Wong. Generalized vector spaces model in information retrieval. Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, 1985: 18-25

[43]Raghavan, V.V., S.M. Wong. A critical analysis of vector space model for information retrieval[J]. Journal of the American Society for information Science, 1986, 37(5): 279-287

[44]Landauer, T.K., P.W. Foltz, D. Laham. An introduction to latent semantic analysis[J]. Discourse processes, 1998, 25(2-3): 259-284

[45]Hofmann, T. Probabilistic latent semantic analysis. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, 1999: 289-296

[46]Ren, S., S. Gao, J. Liao, et al. Improving Cross-Domain Recommendation through Probabilistic Cluster-Level Latent Factor Model. Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015: 4200-4201

[47]Cremonesi, P., A. Tripodi, R. Turrin. Cross-domain recommender systems. Data

Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 2011: 496-503

[48]Koren, Y., R. Bell. Advances in collaborative filtering, in Recommender systems handbook, Springer. 2011, 145-186

[49]Mcauliffe, J.D., D.M. Blei. Supervised topic models. Advances in neural information processing systems, 2008: 121-128

[50]Xin, X., Z. Liu, C. Lin, et al. Cross-domain collaborative filtering with review text. Proceedings of the 24th International Conference on Artificial Intelligence, 2015: 1827-1833

[51]Zhong, E., W. Fan, Q. Yang. Contextual Collaborative Filtering via Hierarchical Matrix Factorization. SDM, 2012: 744-755

[52]Davis, J., M. Goadrich. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning, 2006: 233-240

[53]Willmott, C.J., K. Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance[J]. Climate research, 2005, 30(1): 79

攻读硕士学位期间主要研究成果

[1] Wei S, Xiao L, Zheng X, et al. A Hybrid Movie Recommendation Approach via Social Tags[C]//e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on. IEEE, 2014: 280-285.

致谢

两年半的研究生生涯即将画上句号，自己也即将结束学生时代，进入工作岗位。在这两年半的硕士生涯中，学习到了很多知识技术，也更收获到了来自老师的教诲以及同学的友谊。

首先我要向陈德人老师和郑小林老师表达最真挚的感谢！有幸来到浙江大学，更有幸来到电子商务实验室成为两位导师的学生，陈老师严谨的治学风格，诲人不倦的高尚师德，朴实无华的人格魅力深深的影响了我。郑老师平时的办公环境和学生融合在一起，工作一丝不苟，经常很晚时还在实验室忙碌，而对于学生非常平易近人，在学术科研上寄予指导与帮助。

此外，还要感谢在两年半中所有遇到的同学们，每天一起的欢乐与说笑，度过了快乐的研究生生涯。感谢同级的小伙伴，洪福兴、赵家骏、马国芳，每天一起欢乐的时光，我会深藏心中；感谢陈超超、丁伟峰、扈中凯、魏守贤、林臻师兄，学术与生活上寄予的指导与帮助，我会铭记于心；感谢倪泽明、邓志豪、叶夏菁、许欢、许凌之学长学姐，你们是我的榜样；感谢苏赞文、方崇豪、朱梦莹、王梦晗、黄旭东、王磊、高艳梅等学弟学妹们，我们一起做项目，一起欢乐学习。

感谢我的父母，他们给予了我快乐无忧的成长环境，也造就了我开朗乐观的性格。也感谢我的女友，一直陪伴着我。还有我的室友，愿我们友谊长存。

最后在离开校园之际，向浙江大学表示衷心的感谢，感谢这两年半快乐的学生生活。

肖力涛

2016 年 1 月