浙江大學

本科生毕业论文
开题报告

学生姓名：杨煜溟

学生学号：3130000328

指导教师：郑小林

专　　业：2013 级　计算机科学与技术

学　　院：计算机科学与技术学院

一、题目：结合标签主题的跨域推荐系统研究

二、指导教师对开题报告、外文翻译和文献综述的具体要求：

1. 文献综述要求围绕个性化推荐系统的国内外研究现状进行深入分析，阅读文献20篇以上，形成对推荐系统相关研究的深入理解，分析存在的问题。
2. 外文翻译要求选择与推荐系统相关的经典文献，翻译必须做到语句通顺，语义贴切。
3. 在此基础上，开题报告要提出跨域推荐相应的解决方案，提出可行的技术路线，以及合理的研究计划。

指导教师（签名）：

年　　月　　日

毕业论文开题报告、外文翻译和文献综述考核

导师对开题报告、外文翻译和文献综述评语及成绩评定：

成绩比例	开题报告 占 (20%)	中期报告 占 (10%)	外文翻译 占 (10%)
分值			

导师签字 _____

年　月　日

答辩小组对开题报告、外文翻译和文献综述评语及成绩评定：

成绩比例	开题报告 占 (20%)	文献综述 占 (10%)	外文翻译 占 (10%)
分值			

答辩小组负责人(签名) _____

年　月　日

目　录

“结合标签主题的跨域推荐系统研究”开题报告

1 课题背景

1.1 研究背景

随着互联网的发展，海量信息的复杂性和不均匀性使得信息检索变得困难而耗时，如何处理信息过载的问题成为了一项挑战。个性化推荐系统根据用户过往的行为，分析用户的兴趣模式，自动为用户过滤掉低相关的内容，呈现符合品味的个性化建议，大大降低了用户检索信息的成本。

应用个性化推荐技术需要两个条件，第一个是存在信息过载的情况，第二个是用户在大部分时候没有明确的需求。越来越多的网站成功地引入了推荐系统，广泛利用推荐系统的领域包括电子商务、电影和视频、音乐、社交网络、基于位置的服务、个性化广告等。

推荐系统可追溯到很多相关研究领域，例如认知科学、机器学习和信息检索等[1]。由于其与日俱增的重要性，它在 20 世纪 90 年代发展成一个独立的研究领域。在推荐的过程中，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。

1.2 研究现状

推荐算法的本质是通过一定方式将用户和物品联系起来，常用的方式有利用好友关系、用户的历史兴趣以及用户的注册信息等 [2]。概括地说，推荐系统主要基于两种不同的方法或它们的组合:基于内容的方法和基于协同过滤的方法。

基于内容的方法为每个用户或物品赋予简要的描述，以表示各自独特的性质，这样就可以利用描述为用户匹配合适的物品。这种方式的好处是透明度高，推荐方式直接，而且当有新物品出现时，利用物品的描述即可进行推荐。当然，缺点也很明显，基于内容的策略需要收集额外的信息，而这些信息可能并不容易得到，同时隐私问题也可能阻碍用户提供个人信息 [3]。

另一种策略，不像内容过滤那样需要明确的描述信息，而是通过分析用户的历史行为信息对用户的兴趣进行建模，以获得用户和物品的关联，这种方法被称为协同过滤(Collaborative Filtering)[4]。协同过滤算法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。协同过滤主要的两个领域是基于邻域的方法

和基于潜在特征模型的方法，后者尝试从偏好度矩阵中推断出用户和物品的低维的特征向量映射，这些方法因为具有良好的可扩展性和预测精确性而变得流行。

1.3 存在的挑战

推荐系统需要根据用户的历史行为预测未来的行为和兴趣，因此大量的用户行为数据是实现推荐系统的前提，而在没有大量数据的情况下，如何设计出让用户满意的推荐系统就是冷启动问题，冷启动问题一般分为三类：用户冷启动、物品冷启动、系统冷启动。另外，用户物品的偏好度矩阵通常是非常稀疏的，因为单个用户浏览或使用过的物品只是很小的一部分，这样的稀疏矩阵导致潜在的关联度降低，影响推荐算法对用户兴趣的建模。如何克服冷启动和数据稀疏性问题是目前推荐系统研究领域的热点。

利用内容信息可以缓解数据稀疏性和冷启动问题，一般情况，可以利用向量空间模型将文本表示成关键词向量的形式，但是，对于关键词很少的短文本，向量空间模型的准确性会大大降低。主题分布提供了文档的低维表示，代表性的主题模型有隐式狄利克雷分布（LDA），该模型的假设是文章与词之间是通过主题联系的。

跨域推荐是一个新兴的研究课题，它旨在利用辅助域中的用户反馈来缓解目标域上的稀疏性问题。现有的推荐系统大多是仅针对属于单个域内的用户物品进行预测推荐，因此是在单一域上的建模。事实上，用户在不同域中的偏好之间可能存在依赖性和相关性，因此，在一个域中获得的用户兴趣特征可以在几个其他域中传递和利用，而不是独立地处理每种类型的项目。虽然跨域推荐的效果可能不如在单一域上的推荐准确，但跨域推荐将更加多样化，这可能对提高用户的满意度和参与度有好处[5]。

跨域推荐的关键挑战是在不同域的项目和用户之间发现有用的联系，通常所考虑的域之间看上去是不相关的，例如，音乐与感兴趣的地方，很难找到它们之间的关联 [6]。同时，现有的方法大都要求不同域之间有共享的用户，即存在一些用户在多个域上都有行为数据。然而，更具挑战的是如何在没有共享用户的情况下进行跨域推荐。

2 目标和内容

通过前期的文献调研，我们知道协同过滤克服了基于内容推荐的一些限制，它比内容过滤的技术更加精确，但是却无法解决冷启动问题和数据稀疏性问题。

跨域推荐尝试利用辅助域中的信息来协助目标域上的推荐，为解决协同过滤的冷启动问题和数据稀疏性问题提供了有意义的方向。标签可以作为连接不同域的桥梁，因为不同域中使用的标签词汇之间的通常是重叠的。现有的方法大都要求不同域之间有共享的用户，我们希望找到在没有共享用户的情况下关联多个域的方法，利用跨域推荐提高推荐系统的效果。

基于以上所述，我们目标是研究使用主题建模采集标签中的语义信息，以主题作为不同域之间的桥梁，结合传统的 SVD 协同过滤方法，利用辅助域的信息缓解目标域的冷启动问题和数据稀疏性问题。

在这个背景下，我们的研究大致有如下几项任务：

1. 获取数据。使用 HetRec 2011 中包含的数据集，这些数据集包含标签和社交关系等丰富的信息。
2. 提取标签主题。对每个用户和物品的标签集合，利用主题建模提取文本主题。
3. 矩阵分解建模。得到标签集合的主题后，我们将其加入到 SVD 矩阵分解中，得到用户和物品的潜在向量。
4. 实验和分析。设计实验来评估所提出的预测模型，在跨域的场景下计算模型的预测准确度，并与使用单个域数据的方法进行对比，总结不同方法间各自的优劣。

以上每一步都会用到一些算法和技术，本项目将研究将这些算法和技术整合到跨域推荐中来的可行方法。

3 可行性分析

结合标签的跨域推荐模型 [6] 可以在没有共享用户的情况下，将辅助域和目标域间建立起联系，提高了评分预测的准确性。

隐式狄利克雷分布（LDA）[7] 是经典的概率主题模型，它可以从大量文档集合中发现若干主题，其中主题是关于词项的分布。LDA 属于非监督学习的范畴，给定一个文档语料库，我们可以使用变分 EM 算法来学习主题并根据它们给文档分配主题。

以上的研究成果可以作为本研究的基础部分，同时也表明了提出的目标的可实现性。

4 研究方案和关键技术考虑

4.1 研究方案

主要研究方案是基于真实的标签系统数据集，对其建立跨域推荐模型，提升推荐系统性能，具体目标已在上文阐述。

在具体研究方法方面，首先要查阅相关的文献资料，了解跨域推荐和自然语言处理方面现有的模型和研究进展。通过对前人经验和成果的总结和理解，对这个领域的知识形成大致的轮廓，进一步在现有模型的基础上探索构建可以满足本文目标的模型。之后在符合跨域场景的数据集上，对提出的模型进行测试，并根据评估结果对模型进行改进。

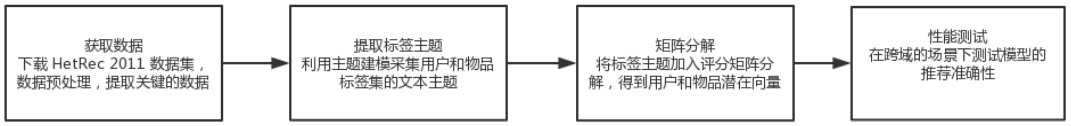

图 4.1 逻辑架构图

本研究的关键算法主要涉及跨域协同过滤和自然语言处理两个方面，下面对研究中的关键技术进行大致的描述。

4.2 关键技术

4.2.1 潜在特征模型

潜在特征模型尝试从偏好度矩阵中推断出用户和物品的低维的特征向量映射，某种意义上，特征向量隐含了用户和物品在多个维度上的性质。在该模型中，用户对物品的预测偏好度是特征向量的线性结合。例如，每一个物品 i 与向量 $q_i \in R^f$ 相关联，每一个用户 u 与向量 $p_i \in R^f$ 相关联，它们的内积 $q_i^T p_u$ 表现了用户 u 对物品 i 在 f 个特征上的总体偏好度。因此评分的估计由如下式子给出：

$$\hat{r_{ui}} = q_i^T p_u.$$

这种方法最主要的挑战是如何将每一个用户和物品映射到特征向量 $q_i, p_u \in R^f$ ，在完成了映射之后，推荐系统将很容易利用上面的公式预测用户对物品的评分。潜在特征向量映射的实现通常是基于矩阵分解的，这些方法因为具有良好的可扩展性和预测精确性而变得流行。

4.2.2 奇异值分解

奇异值分解（SVD）[8] 是一种最基本的矩阵分解方式，它的计算方式是使得到的矩阵与原始矩阵对应项的平方和误差最小。因为大多数的评分矩阵都是相当稀疏的，所以它只关注这些很少的值会导致过拟合问题。早期通过填补矩阵中缺失的评级使矩阵变得稠密，但是随着可见项的增加，计算量可能难以承受，另外，不准确的填充会严重影响预测的效果。可以通过引入正则项缓解过拟合的问题，为了得到特征向量，系统最小化在已知评分上的正则平方误差：

$$\min_{q^*,p^*} \sum_{(u,i)\in} \left(r_{u,i} - q_i^T p_u\right)^2 + \lambda(||q_i||^2 + ||p_u||^2),$$

这里，κ 是训练集中所有已知评级的用户物品对 (u,i) 的集合，系统通过拟合之前观测的样本来学习模型的参数，而我们的目标是预测未知的评分，所以应该通过正则化参数来避免过度拟合已知的项，常数 λ 用于控制正则化的程度。可以通过随机梯度下降或迭代最小二乘的方法最小化上面的式子。

随机梯度下降算法（stochastic gradient descent）最优化理论里最基础的优化算法，它首先通过求参数的偏导数找到函数的最速下降方向，然后通过不断迭代优化参数直至收敛。上面定义的损失函数里有两组参数 p_u 和 q_i ，对它们分别求偏导数，然后梯度相反的方向以 步长调整参数，可以得到如下的迭代公式：

$$q_i \leftarrow q_i + \cdot(e_{ui} \cdot p_u - \lambda \cdot q_i)$$

$$p_u \leftarrow p_u + \cdot(e_{ui} \cdot q_i - \lambda \cdot p_u)$$

4.2.3 概率主题模型

主题建模算法用于从大量文档集合中发现一组主题，其中主题是关于词项的分布，主题模型提供了文档的低维表示 [9]。最常见的主题模型是隐式狄利克雷分布（LDA）[7]，假设有 K 个主题 $\beta_{1:k}$，每一个是在固定词典上的分布。LDA 生成文档的大致流程如下：对于语料库中的每一篇文档 w_{jn} ：

1. 从狄利克雷分布中选取主题分布 $\theta_j \sim Dirichlet(\alpha)$.

2. 对于文档中的每一个词 n ：

(a) 选取主题 $z_{jn} \sim Mult(\theta_j)$.

(b) 选取单词 $w_{jn} \sim Mult(\beta_{z_{jn}})$.

这个过程说明了文档中的每个词是如何从主题的集合中选取出来的：主题分布是文档特有的，但是主题的集合是整个语料库共享的。

LDA 属于非监督学习的范畴，给定一个文档语料库，我们可以使用变分 EM 算法来学习主题并根据它们给文档分配主题。此外，给定一个新的文档，我们可以使用变分推理来确定其内容的主题 [7]。

5 预期研究结果

本研究希望提出一种新的跨域推荐模型，以文本主题作为域之间的桥梁，结合传统的协同过滤方法，将辅助域的信息迁移至目标域，缓解单个域推荐时的冷启动问题和稀疏性问题，以求获得较高的推荐准确性。

6 进度计划

根据之前所述的研究方法和预期结果，将论文的进度计划安排如下：

时间	进度安排
2016.7 - 2016.9	了解推荐系统领域的研究方向，并学习机器学习的相关知识
2016.9 - 2016.12	阅读相关文献资料，充分理解现有的研究成果，并撰写文献综述
2017.2.1 - 2017.2.28	提出初步的研究方案，与导师进行讨论，做出补充和改进
2017.3.1 - 2017.3.10	根据研究方案确定初步的模型
2017.3.11 - 2017.3.25	获取实验数据，分析数据并进行预处理
2017.3.26 - 2017.4.10	实现论文中的关键算法，并在数据集上测试效果
2017.4.11 - 2017.4.20	对比不同算法，并对各项指标进行分析
2017.4.21 - 2017.4.30	对研究结果进行归纳，整理实验数据，完成论文初稿
2017.5.1 - 2017.5.15	完善和修改，并确定论文终稿

参 考 文 献

[1] 肖力涛. 基于隐式因子和隐式主题的跨域推荐算法研究. Master's thesis, 浙江大学, 2016.

[2] 项亮. *推荐系统实践*. 人民邮电出版社, 2012.

[3] Y Koren, R Bell, and C Volinsky. Matrix factorization techniques for recommender systems. Computer, 42(8):30–37, 2009.

[4] David Goldberg. Using collaborative filtering to weave an information tapestry. Communications of the Acm, 35(12):61–70, 1992.

[5] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. Cross-domain recommender systems: A survey of the state of the art. In Spanish Conference on Information Retrieval, 2012.

[6] Yue Shi, Martha Larson, and Alan Hanjalic. Tags as bridges between domains: Improving recommendation with tag-induced cross-domain collaborative filtering. In International Conference on User Modeling, Adaptation, and Personalization, pages 305–316. Springer, 2011.

[7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[8] Arkadiusz Paterek. Improving regularized singular value decomposition for collaborative filtering. In Proceedings of KDD cup and workshop, volume 2007, pages 5–8, 2007.

[9] Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves: How humans interpret topic models. Advances in Neural Information Processing Systems, 32:288–296, 2009.