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的产品和内容建议，大大降低了用户检索信息的成本。

个性化推荐的成功应用需要两个条件，第一个是存在信息过载的情况，第二个是用户大部分时候没有明确的需求。在互联网的各类网站中都可以看到推荐系统的应用，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[8]。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。

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、用户的历史兴趣记录以及用户的注册信息等 [9]。

概括地说，推荐系统主要基于两种不同的策略或其组合：基于内容的过滤方法和协同过滤方法。基于内容的过滤方法为每个用户或物品创建描述以表征其性质，这样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 [5]。

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建模，以获得新的用户和物品的关联，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。协同过滤的两个主要领域是基于邻域的方法

和潜在因素模型，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，而对于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，这时，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，它旨在利用相关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[3]。

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，使得难以找到它们之间的关联 [7]。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，但是却无法存在的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用户反馈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 [7] 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）[1] 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。

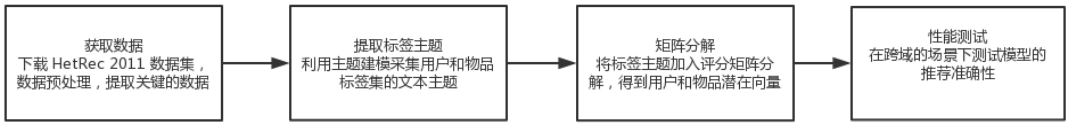


图 4.1 逻辑架构图

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）[6] 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 [2]。最常见的主题模型是隐式狄利克雷分布（LDA）[1]，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 [1]。

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

[2] 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.

[3] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. Cross-domain recommender systems: A survey of the state of the art. 2012.

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

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

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

[7] 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.

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

[9] 项亮. *推荐系统实践*. 人民邮电出版社, 2012.

本科毕业论文外文翻译

文献原文：

Chen C, Zheng X, Wang Y, et al. Capturing Semantic Correlation for Item Recommendation in Tagging Systems[C]//AAAI. 2016: 108-114.

利用标签的语义关联性进行物品推荐

摘要 标签系统的普及对于提升物品推荐的效果是一个很好的机会。虽然现有的方法对标签使用主题建模来挖掘物品的语义信息，但是他们忽略了一个重要的性质，标签是用户和物品之间连接的桥梁。因此，这些方法不能处理无共同评分项的数据稀疏性问题（DS-WO-CRI），从而限制了它们的推荐性能。为了解决这个问题，我们提出了一种新型的基于标签和评分的协同过滤推荐模型，首先使用主题建模依次挖掘每个用户和每个物品的语义信息，然后将这些语义信息纳入矩阵分解，同时捕获标签和评分在用户和物品间的桥接特性。因此，我们的模型捕获了用户和物品间的语义关联，极大的提高了推荐性能，尤其是在 DS-WO-CRI 的情况下。在两个流行的真实数据集上的实验表明，我们提出的模型在准确率和召回率上显著优于传统的协同过滤方法、最先进的基于社交关系的协同过滤和基于主题模型的的协同过滤方法，它是解决 DS-WO-CRI 问题的有效方法。

1 引言

近些年来，诸如 Delicious（社交书签），Last.fm（社交音乐），Flickr（照片分享）和 YouTube（视频分享）这些标记系统为用户提供了高效的方式来组织、管理、共享并搜索各种项目。例如，一个人在 Last.fm 听 Lady Gaga 的音乐时，他可以将她标记为“流行的”和“女歌手”。这些连同评分行为一起出现的标签很有价值，强烈建议使用这样的信息来提供个性化推荐 [21] 。

标签系统的普及促进了推荐系统的发展，尤其是标签系统中的协同过滤方法。目前为止，标签系统中主要有两种类型的协同过滤：标签推荐 [19] 的目的是为物品推荐合适的标签，另一种是基于标签的物品推荐 [22, 23]，它关注利用标签和评分等信息为目标用户推荐相似的用户或物品。

目前，一个研究的趋势是在协同过滤中使用主题模型来处理标签信息

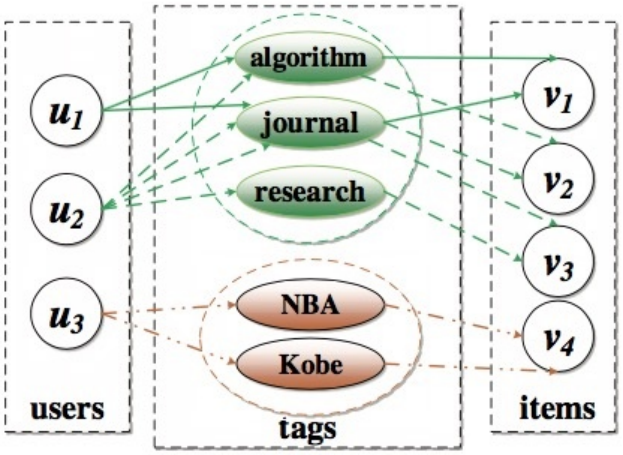


图 1.1　一个标签系统的例子。每个标签都对应了一个评分，为简洁起见省略了评分。

[1, 18, 15, 19, 6]。例如，Wang 和 Blei [18] 提出了一个协同主题回归（CTR）模型，可用于基于标签的物品推荐。Chen 等人 [6] 提出的将 CTR 与社交矩阵分解 [13] 结合的推荐方法获得了更好的预测效果。然而，这些方法只是将标签与用户和物品相关联，而忽略了标签重要性质，标签连接了用户和物品，这与评分的作用是一样的。但是标签可以反应用户和物品间的语义关联性，而评分没有这样的能力。

当一个用户为一些物品赋予了标签，那这些标签就反映了用户对物品的偏好。一个标签越频繁被一个用户使用，越可能表示这个用户喜欢这个标签所指代一类的物品 [21]。类似的，如果一个标签被越多的用户赋予一个物品，那么很可能这个物品匹配这个标签。因此，标签同时包含用户和物品的语义信息，而不仅仅是单独的用户或物品。

一个针对性的例子　图 1.1 描述了一个标签系统的例子，这个系统包含三个用户（u_1, u_2, u_3），四个物品（v_1, v_2, v_3, v_4）和五个标签（“algorithm”, “journal”, “research”, “NBA”, “Kobe”）。

这个例子中，用户 u_1 标注了物品 v_1，用户 u_2 标注了物品 v_2 和 v_3，因此，用户 u_1 和用户 u_2 之间没有共同评分项。我们定义这种情况为 * 无共同评分项的稀疏性问题（DS-WO-CRI），DS-WO-CRI 是标准的数据稀疏性问题（在所有的用户物品对中只有很少比例的已知项）的一个典型子集。在 DS-WO-CRI 的情况下，已有的协同

过滤方法，例如，PMF 和 CTR 模型，都无法将物品 v_1 推荐给用户 u_2 ，因为这些方法无法捕获它们之间的任何联系。然而，一个好的推荐系统应该能够将物品 v_2 和 v_3 推荐给用户 u_1 并且将物品 v_1 推荐给用户 u_2，因为在这个例子中，用户 u_1 和 u_2 很可能是算法方面的研究者，而物品 v_1, v_2 和 v_3 可能跟算法有关。

现有的研究表明，用户对一个物品的评分或标注等动作就可以表明用户喜好该物品，而不需考虑评分的等级 [11, 12]。换句话说，用户通过标记和评分隐式的表达了他的偏好 [11]。因此，一个用户和他所标注和评分的物品趋向于具有相似的潜在特征，我们在本文中将其定义为隐式偏好（implicit preference）。在上面的例子中，用户 u_2 将物品 v_3 标注为“journal”和“research”，这在表明用户 u_2 可能是一名研究者同时，也指出物品 v_3 可能是一篇学术期刊或其他相关的东西。因此，在语义上，用户 u_2 和物品 v_3 的潜在特征应该具有某种程度的相似。

然而，现有的方法无法捕获用户和物品之间的语义关系，因此它们的推荐性能被局限了，尤其是在 DS-WO-CRI 的情况下。

我们的提议 为了解决上面提到的 DS-WO-CRI 问题，我们在这篇文章中提出了一种新型的协同过滤系统。我们首先利用主题模型依次为每一个用户和每一个物品挖掘标签的语义信息，然后将这些语义信息纳入矩阵分解，同时捕获标签和评分在用户和物品间的桥接特性（即，隐式偏好）。因此，我们的模型可以捕获用户和项目间的语义关联，并将具有相似语义信息的物品推荐给用户，即便是在 DS-WO-CRI 的情况下。

贡献 我们的工作主要有以下贡献：（1）我们首先指出标签的重要特性，即它们做为桥梁将用户和物品连接起来，概述了用户和物品之间的语义联系，然后我们说明了利用此特性可以帮助解决 DS-WO-CRI 问题。据我们所知，这是首个针对这个问题的研究。（2）我们提出了一种新型的基于标签和评分的协同过滤模型，它可以捕获用户和物品之间的语义联系，因此可以大大地提升推荐性能，尤其是在 DS-WO-CRI 的情况下。我们还提出了基于坐标上升的参数学习方法。据我们所知，这项研究是文献中对捕获用户和物品语义关联的首次尝试。（3）在两个流行的真实世界的数据

集上的实验表明，我们提出的模型在精确率和召回率方面都显著优于最先进的方法。实验还表明，我们的模型是一个解决 DS-WO-CRI 问题的有效方法。

2 相关工作

在本节中，我们将分三组来回顾已有的物品推荐方法，其中包括：传统的协同过滤方法、基于主题模型的协同过滤方法、以及基于社交关系的协同过滤方法。

基于已有的研究 [17]，传统的协同过滤方法仅利用用户物品的评分进行推荐，主要分为两种类型：基于内存的协同过滤 [8] 和基于模型的协同过滤 [12, 11, 22, 20]，这两种方法都可以用于标签系统的推荐。

传统的协同过滤方法无法借助文本内容的信息，因此，一些混合模型被提出来，它们结合基于内容的方法和协同过滤方法进行推荐 [14]。但是，这些方法简单的将内容表示为词向量的形式，因而无法发掘它们的语义信息。为了利用内容所提供的语义信息，研究者利用主题模型提高推荐效果，Agarwal 等人提出了 fLDA 模型 [1]，该方法通过将 LDA 中学习到的先验信息加入物品向量，结合了 RLFM[1] 和隐式狄利克雷分布（LDA）。RLFM 和 fLDA 都能纳入额外的原信息，例如用户年龄和物品类别，然而，这些附加元特征信息不在本文的范围之内。稍后的，Wang 等人 [18] 将概率矩阵分解 [16] 与 LDA [4] 相结合，提出了协同主题回归模型。在 [18] 中证明了在相似的情况下 CTR 的性能要优于 fLDA，因为 fLDA 基本忽略了其他用户的评分。

此外，用户之间和物品之间的社会信息对于提升推荐的性能是有效的 [7]。首先，用户的社会信息被纳入常规的协同过滤模型 [10]，例如，Ma 等人提出 Soreg 来约束具有联系的用户的潜在因子之间的差异性。之后，相邻用户和物品的社会信息被引入了基于主题模型的协同过滤中以进一步改善推荐性能，例如，Purushotham、Liu 和 Kuo [15] 以及 Chen 等人 [6] 提出了两种模型（CTR-SMF 和 CTR-SMF2），将用户社交关系网络纳入 CTR，以进一步提高项目推荐性能。Wang、Chen 和 Li [6] 提出了一个将物品的社会关系引入 CTR 的模型，以提高标签系统中的标签推荐性能。

3 提出的模型——TRCF

在本节中，我们提出一个新的基于标签和评分的协同过滤方法（TRCF）。我们首先正式确定基于标签的物品推荐问题并定义一些符号。然后，我们提出 TRCF ，这是一个分层的贝叶斯模型。最后，我们提出了基于坐标上升的参数学习方法。

3.1 初步定义

假定，我们有一个用户的集合 $U = \{u_1, \ldots, u_I\}$，这些用户用一组标签 $T = \{t_1, \ldots, t_N\}$ 标记了的一组物品 $V = \{v_1, \ldots, v_J\}$，以及评分的集合 $R = \{r_1, \ldots, r_O\}$，其中，I 、J、N 和 O 依次代表了用户、物品、标签和评分的数目。每一个用户-物品-标签-评分（U-I-T-R）的可观察数据是一个四元组 $\square u_i, v_j, T_{ij}, R_{ij} \square$，其中 $u_i \in U$，$v_j \in V$，T_{ij} 是用户 u_i 给予物品 v_j 的标签集合，并且 $T_{ij} \in T$ ，R_{ij} 是用户 u_i 给予物品 v_j 的评分，评分基于用户对该物品的喜好程度并在同时标注它。然而，用户-物品（U-I）的评分集合 R 一般是整数集合，例如，*MovieLens* 使用 [1,5] 范围内的评分。$U \in R^{K*I}$ 表示潜在的用户特征矩阵，其中列向量 U_i 表示属于用户 u_i 的 K 维潜在特征向量。$V \in T^{K*J}$ 表示潜在的物品特征矩阵，其中列向量 V_j 表示属于物品 v_i 的 K 维潜在特征向量。

对于基于标签和评分的物品推荐中，给定现有的四元组 U-I-T-R ，我们的目标是预测用户 u_i 对物品 v_j 的未知的评分。

3.2 基于标签和评分的协同过滤

TRCF 是一个新型的分层的贝叶斯模型，图 3.1 展示了它的图模型，其中 N_u 和 N_v 依次表示用户 u_i 和物品 v_j 的标签数目。我们首先将每个用户和物品的标签依次分组，然后使用隐式狄利克雷分布依次对每个用户和每个物品的标签集合进行语义挖掘（图 3.1 中以红色绘制）。最后将这些语义信息纳入矩阵分解用于分解评分信息（图 3.1 中以紫色绘制）以及捕获标签和评分所提供的隐式偏好（图 3.1 中以蓝色绘制）。

TRCF 同时在用户和物品两个方面上执行 LDA ，因此可以同时捕获用户和物品的语义信息，而不仅仅像现有的工作那样只捕获物品的语义信息。另外，在 TRCF 中，如果一个用户和一个物品通过标签或评分相关联，那么他们的隐式特征会在某些程度上比较相似，这被称为隐式偏好。相比之下，现有的基于主题建模的 CF 方法，例如，CTR、CTR-SMF 和 CTR-SMF2，都是假设用户和物品是独立的，并且忽略标签和评分在用户和项目之间的桥接作用。因此，TRCF 可以捕获用户和物品之间的语义关联，并且能够处理 DS-WO-CRI 问题。假定每个用户和每个物品都有 K 个主题，TRCF 的执行过程如下：

1. **挖掘用户标签的语义信息。** 对于每个用户 u_i ：

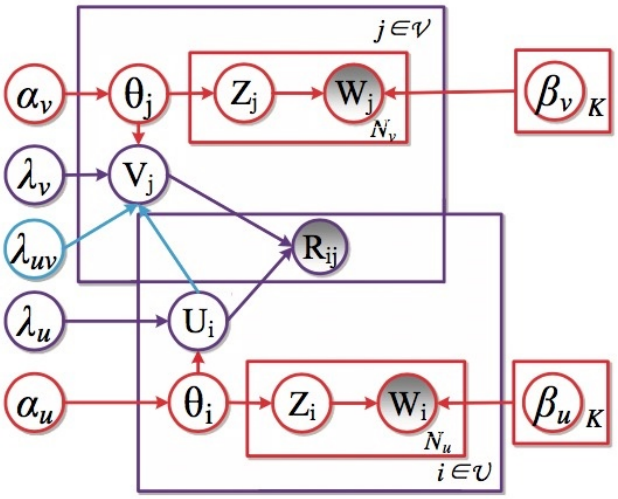


图 3.1　TRCF 的图模型。LDA 部分以红色绘制，隐式偏好部分以蓝色绘制，PMF 部分以紫色绘制。

(a) 选取主题分布 $\theta_i \sim Dirichlet(\alpha_u)$ ；

(b) 选取用户的潜在向量 $U_i \sim \mathcal{N}(\theta_i, \lambda_u^{-1} I_k)$ ；

(c) 对于用户 u_i 的每一个标签 w_{in_u} ：

i. 选取主题 $z_{in_u} \sim Mult(\theta_i)$ ；

ii. 选取标签 $w_{in_u} \sim Mult(\beta_{z_{in_u}})$ ；

2. **挖掘物品标签的语义信息，并且采集用户和物品间的隐式偏好。**对于每一个物品 v_j ：

(a) 选取主题分布 $\theta_j \sim Dirichlet(\alpha_v)$ ；

(b) 选取物品的潜在向量 $V_j \sim \mathcal{N}(\theta_j, \lambda_v^{-1} I_k) \times \prod_i I_{ij}^R \mathcal{N}(U_i, \lambda_{uv}^{-1} I_K)$ ；

(c) 对于物品 v_j 的每一个标签 w_{jn_v} ：

i. 选取主题 $z_{jn_v} \sim Mult(\theta_j)$ ；

ii. 选取标签 $w_{jn_v} \sim Mult(\beta_{z_{jn_v}})$ ；

3. **获取评分。**对于每一个用户-物品对 (i, j) ：

$$R_{ij} \sim (U_i^T V_j, c_{ij}^{-1}).$$

在上面的生成过程中，$\mathcal{N} \sim (x|\mu, \sigma^2)$ 是期望为 μ 方差为 σ^2 的高斯分布，I_K 是一个 K 行 K 列的单位矩阵。I_{ij}^R 是一个指示函数，如果用户 u_i 为物品 v_j 打分了，它的值为 1，否则为 0。C 是评分置信度矩阵，其中的项 c_{ij} 表示评分的置信度。更多细节请参考 [18]。

参数 λ_u 平衡了用户语义信息提供的标签和评分信息对模型性能的影响。类似地，参数 λ_v 平衡了由物品语义信息提供的标签和评分信息对推荐性能的影响。参数 λ_{uv} 平衡隐式偏好对模型性能的贡献，即通过评级和标签链接的用户和项目之前潜在特征相似度的程度。

观察到的评分的条件分布可以被形式化为：

$$p(R|U, V, C) = \prod_i \prod_j \mathcal{N}(R_{ij}|U_i^T V_j, c_{ij}).$$

用户和物品的潜在向量 U_i 和 V_j 生成的方式与 CTR 相似，它们可以被形式化为：

$$\begin{aligned} p(U|\lambda_u) &\sim \prod_i \mathcal{N}(\theta_i, \lambda_u^{-1} I_K), \\ p(V|U, \lambda_v, \lambda_{uv}) &\sim \prod_j \mathcal{N}(\theta_j, \lambda_v^{-1} I_K) \times \prod_i I_{ij}^R \mathcal{N}(U_i, \lambda_{uv}^{-1} I_K). \end{aligned} \tag{3.1}$$

给定了 U-I-T-R 信息，通过使用贝叶斯推理，我们可以得到 TRCF 的潜在特征向量的后验概率的如下式子：

$$p(U, V|R, C, \lambda_u, \lambda_v, \lambda_{uv}) \propto p(R|U, V, C) p(U|\lambda_u) p(V|U, \lambda_v, \lambda_{uv}). \tag{1}$$

3.3 TRCF 的参数学习

给定主体参数 β_u 和 β_v ，直接计算 $U_i, V_j, \theta_i, \theta_j$ 的完整后验是困难的。我们使用坐标上升的方法来学习最大后验概率。使等式 (1) 中具有固定超参数的两个潜在特征的后验最大化等价于，在给定 λ_u, λ_v 的条件下，使如下的 $U, V, \theta_{1:I}, \theta_{1:J}$ 的对数似

然函数最大:

$$\begin{aligned} L = & -\frac{\lambda_u}{2}\sum_i (U_i-\theta_i)^T(U_i-\theta_i) \\ & -\frac{\lambda_v}{2}\sum_j (V_j-\theta_j)^T(V_j-\theta_j) \\ & -\sum_{ij}\frac{c_{ij}}{2}(R_{ij}-U_i^TV_j)^2 \\ & +\sum_i\sum_{n_u} log\left(\sum_k \theta_{ik}\beta_{k,w_{in_u}}\right) \\ & +\sum_j\sum_{n_v} log\left(\sum_k \theta_{jk}\beta_{k,w_{jn_v}}\right) \\ & --\frac{\lambda_{uv}}{2}I_{ij}^R\sum_{ij}(U_i-V_j)^T(U_i-V_j). \end{aligned} \tag{2}$$

我们省略一个常数并设置狄利克雷先验 $\alpha_u = \alpha_v = 1$ 。这个函数可以通过使用坐标上升来优化,也就是说,我们固定 β_u 和 β_v ,并迭代优化 MF 变量 U_i, V_j 和主题分布 θ_i, θ_j 。具体来说,我们首先根据 θ_i, θ_j 当前的估计值更新 U_i 和 V_j ,我们计算等式 (2) 中 L 在 U_i□V_j 上的导数,并且将它设置为 0 :

$$\frac{\partial L}{\partial U_i} = 0, \frac{\partial L}{\partial V_j} = 0. \tag{3}$$

解决上述的公式得到参数更新的式子:

$$\begin{aligned} U &\leftarrow \left(VC_iV^T + \lambda_u I_K + \lambda_{uv}\sum_i I_{ij}^R I_K\right)^{-1}(VC_iR_i + \lambda_u\theta_i + \lambda_{uv}\sum_j I_{ij}^R V_j), \\ V &\leftarrow \left(UC_jU^T + \lambda_v I_K + \lambda_{uv}\sum_i I_{ij}^R I_K\right)^{-1}(UC_jR_j + \lambda_v\theta_j + \lambda_{uv}\sum_i I_{ij}^R U_i). \end{aligned} \tag{4}$$

其中对于每一个用户 u_i ,C_i 是一个对角矩阵,它的对角元素是 $c_{ij}, j = 1, \ldots, J$,并且 $R_i = R_{ij}{}_{j=1}^{J}$ 。对于物品 v_j ,C_j 和 R_j 的定义是相似的。

公式 (4) 显示了参数 λ_u、λ_v 和 λ_{uv} 是如何影响用户和物品的潜在特征的。越大的 λ_u 会导致用户的潜在特征越依赖于用户标签,而不是评分信息。类似的,较大的 λ_v 表示物品的潜在特征来自物品标签的比例更大,而不是评分信息。此外,更大的 λ_{uv} 意味着更强的约束,即通过标签和评分链接的用户和项目应当具有类似的潜在特征,即隐式偏好。从公式 (4) 可以看出,概率矩阵分解(PMF)和协同主题回归(CTR)都是 TRCF 的特殊形式。

接下来,给定当前的 MF 变量 U_i, V_j ,我们更新主题分布参数 θ_i 和 θ_j 。对于 θ_i ,我们先定义 $q(z_{in_u} = k) = \phi_{in_uk}$,然后分离包含 θ_i 的用户并应用 Jensen 不等式:

$$\begin{aligned} L(\theta_i) \geq & -\frac{\lambda_u}{2}(U_i - \theta_i)^T(U_i - \theta_i) \\ & + \sum_{n_u}\sum_{k} \phi_{in_uk}(log\theta_{ik}\beta_{k,w_{in_u}} - log\phi_{in_uk}) \\ & = L(\theta_i, \phi_i). \end{aligned} \tag{3.2}$$

其中,$\phi_i = \phi_{in_uk}{}_{n_u=1,k=1}^{N_u \times K}$ 。显然 $L(\theta_i, \phi_i)$ 是 $L(\theta_i)$ 的严格下界,因此我们可以使用映射梯度 [3] 来优化 θ_i 。最优的 ϕ_{in_uk} 正比于 $\theta_{ik}\beta_{k,w_{in_u}}$ 。对于 θ_j ,更新的规则是相似的。

对于 β_u ,我们像 LDA 那样为主题执行 M 步的更新。

$$\beta_{kw_i} \propto \sum_{i}\sum_{n_u} \phi_{in_uk} 1[w_{in_u} = w].$$

对于 β_v ,它的更新策略是相似的。当参数 $U^*, V^*, \theta_{1:I}{}^*, \theta_{1:J}{}^*, \beta_u{}^*, \beta_v{}^*$ 的最优值学习完成后,我们的模型就可以进行评分预测了:

$$R_{ij}{}^* \approx (U_i^*)^T V_j^*.$$

4 实验和分析

在本节中,我们介绍了对两个流行的现实世界数据集进行的实验,目的是回答如下的问题:(1)我们的模型相比现有的最先进的方法有什么改进? (2)我们的方法如何处理 DS-WO-CRI 问题? (3)参数 λ_u ,λ_v 和 λ_{uv} 如何影响 TRCF 的性能?

4.1 数据集

我们在实验中使用了两个现实世界的数据集: hetrec2011-delicious-2k (Delicious) 和 hetrec2011-lastfm-2k (Lastfm)[5]。这两个数据集已广泛用于标签系统的实验 [2],它们在表 4.1 中描述。

对于每个数据集,如果用户已将该项目设为书签(或收听),则我们认为该项目的用户评分为 1 ,否则,该项目的用户评分为 0 。

在我们的实验中,我们将每个数据集分为三个部分——训练数据集(80%),留出验证数据集(10%)和测试数据集(10%)。我们在训练数据集上训练我们的模型,在验证数据集上获得最佳参数,并在测试数据集上评估我们的模型。

表 4.1 数据集描述

数据集	用户	物品	标签	用户-标签-物品	评分
Delicious	1867	69226	53388	437593	104799
LastFm	1892	17632	11946	186479	92834

4.2 比较与评估

如相关工作中所述，存在许多种推荐方法，例如，基于存储的方法和混合方法。这里，我们将所提出的 TRCF 与以下三种现有的方法进行比较，即常规的协同过滤方法，基于社会关系的 CF 方法和基于主题建模的方法：

SVD ++ [11] 是一种经典的协同过滤方法，该方法仅使用 U-I 评分信息。

Soreg [13] 是基于社会关系的 CF 方法，其使用 U-I 评分和用户社交信息。

CTR [18] 是一种最先进的基于主题建模的 CF 方法，其使用类似于 TRCF 中使用的四元组 U-I-T-R 的信息。

CTR-SMF [15] 结合了用户社交矩阵分解和 CTR 。它包含除了在 TRCF 中使用的四元组 U-I-T-R 之外的附加用户社交信息。

CTR-SMF2 [6] 改进了 CTR-SMF ，它还将用户社交信息引入了 TRCF 中的 U-I-T-R 四元组。

精确率和召回率已被广泛用作评估推荐效果的指标 [9]。因此，我们使用 Precision 和 Recall 来评估推荐性能，并且计算召回率的方式也用于 CTR，CTR-SMF 和 CTR-SMF2。对于每个用户，Precision 和 Recall 定义如下：

$$\begin{aligned} Precision@M &= \frac{\text{\#Top M 中用户喜欢的物品}}{M}, \\ Recall@M &= \frac{\text{\#Top M 中用户喜欢的物品}}{\text{\#用户喜欢的所有物品}}, \end{aligned} \tag{4.1}$$

其中 M 是推荐列表中的物品数。我们计算测试数据集中所有项的精确率和召回率的平均值作为最终结果。

4.3 性能对比和分析

在对比不同模型时，我们使用在 CTR-SMF [15] 中设置的 SVD++、CTR 和 CTR-SMF 的最佳参数，它们使用相同的数据集。对于 Soreg、CTR-SMF2 和我们的模型，我们使用网格搜索来获得最佳参数。

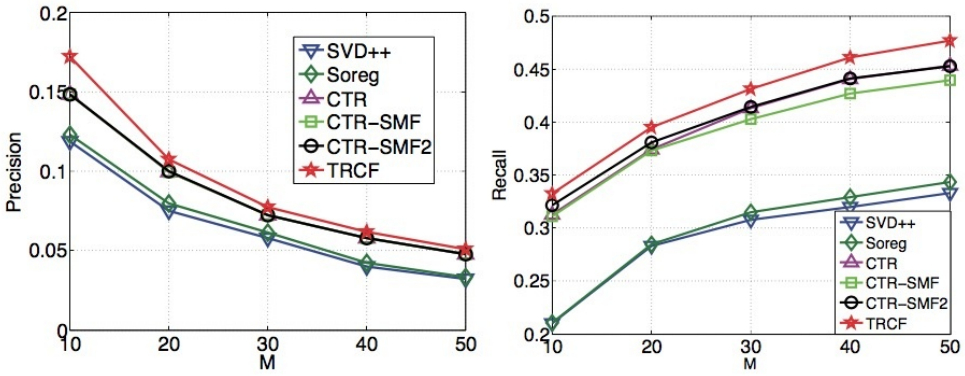


图 4.1　在 Delicious 数据集下设置不同的 M 值，精确率和召回率的对比，以及每个方法的最优参数

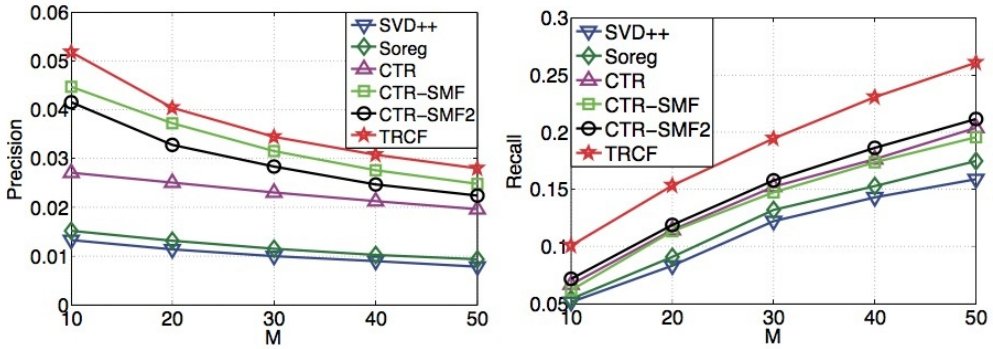


图 4.2　在 LastFm 数据集下设置不同的 M 值，精确率和召回率的对比，以及每个方法的最优参数

结果： 图 4.1 和图 4.2 显示了各个推荐方法在 Delicious 数据集和 Lastfm 数据集上的总体性能，其中我们设置 $M = 10, 20, 30, 40, 50$ 并将每个方法的参数固定为最佳值。结果表明，传统的 CF 方法（SVD ++）和基于社交关系的 CF 方法（即 Soreg）具有相似的性能。三个基于主题建模的 CF 方法（即 CTR、CTR-SMF 和 CTR-SMF2）具有类似的性能，并且显著优于 SVD++ 和 Soreg，这表明了标签信息在推荐中的重要性。

我们提出的 TRCF 方法在这两个数据集上显著优于 SVD++、Soreg、CTR、CTR-SMF 和 CTR-SMF2。具体在平均值上，在 Delicious 数据集上，TRCF 在精确度方面将 SVD++、Soreg、CTR、CTR-SMF 和 CTR-SMF2 提升了 46.75%、39.74%、8.62%、8.78%和 8.65%，并且在召回率方面分别提高了 44.99%、42.40%、5.23%、7.27%和 4.21%。在 Lastfm 数据集上，TRCF 在精确度方面将 SVD++、Soreg、CTR、CTR-SMF 和 CTR-SMF2 提升了 259.80%、210.27%、57.96%、11.64%和 23.85%，并且在召回率方面分别提高了 73.03%、60.39%、34.18%、39.04%和 28.12%。

分析和总结： 以上的对比表明了我们提出模型的有效性，它捕获了用户和物品之间的语义关联。实验结果表明，尽管用户的社交信息可以改善推荐性能，但是改善推荐性能的更有效的方式是考虑用户和物品的语义相关性。

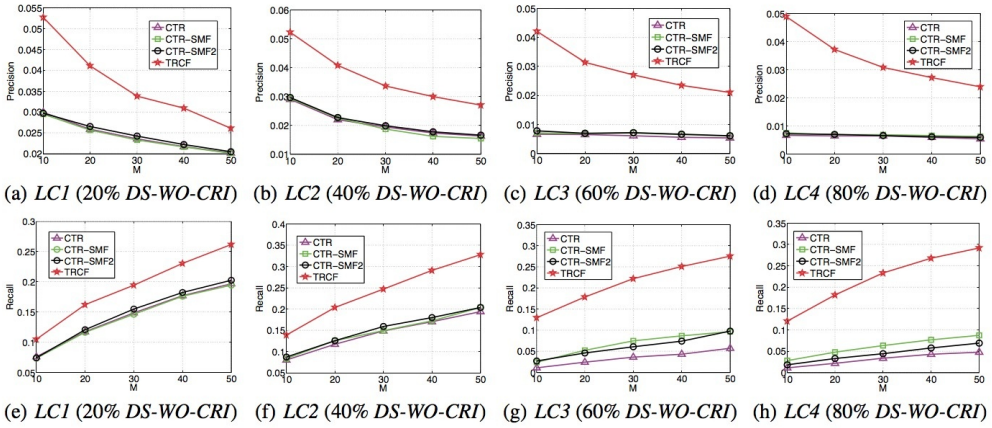
4.4 DS-WO-CRI 实验

所有四种基于主题建模的 CF 方法（包括 TRCF）都可以通过采集物品的语义信息来提高推荐性能。为了研究它们处理 DS-WO-CRI 问题的能力，我们进行以下实验。我们首先依据 DS-WO-CRI 的程度将 LastFm 数据集分为四个子数据集，每个数据集在表 2 中描述了。DS-WO-CRI 的程度定义如下：

$$x\% = \frac{\#\text{没有共同评分物品的用户数目}}{\#\text{总的用户数目}}.$$

然后我们对每个子数据集进行对照实验。图 4.3 显示，我们的模型在不同的 DS-WO-CRI 程度下都达到最佳性能。我们的模型相对于其他三个主题建模方法的在 LC1、LC2、LC3、LC4 上精确率的平均改进分别为 49.67%、75.81%、345.77%和 428.97%，召回率的改进分别为 33.94%、65.00%、383.63 %和 458.00%。实验结果表明，DS-WO-CRI 的程度越严重，我们的模型相对其它模型的优势越明显。

总结： DS-WO-CRI 实验表明了我们提出的模型在处理该问题时的有效性：越



(a) *LC1* (20% *DS-WO-CRI*) (b) *LC2* (40% *DS-WO-CRI*) (c) *LC3* (60% *DS-WO-CRI*) (d) *LC4* (80% *DS-WO-CRI*)

(e) *LC1* (20% *DS-WO-CRI*) (f) *LC2* (40% *DS-WO-CRI*) (g) *LC3* (60% *DS-WO-CRI*) (h) *LC4* (80% *DS-WO-CRI*)

图 4.3　在每个 DS-WO-CRI 子数据集上设置每种方法的最优参数，精确率和召回率对比。

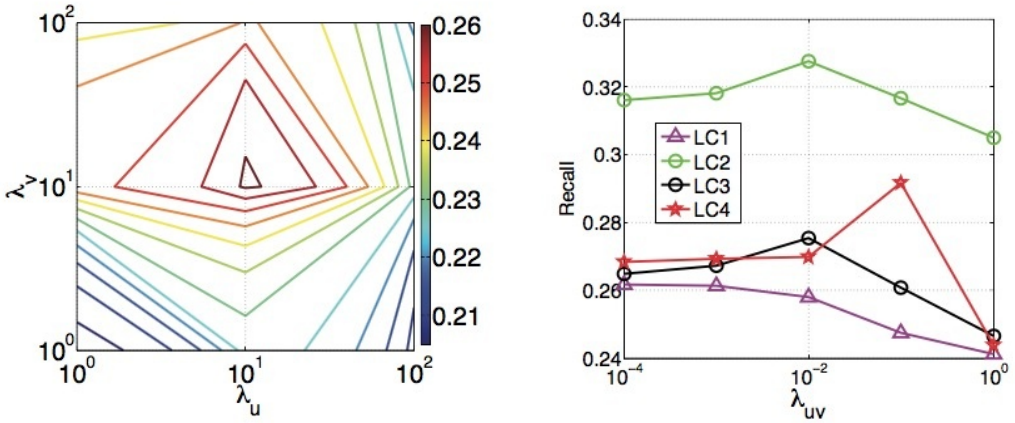
严重的 DS-WO-CRI 情况，我们的模型相比其它模型的优势越明显。这是由于我们的方法具有采集用户和物品间标签语义关联的能力。

4.5　参数影响

图 4.4(a) 显示了在 LastFm 数据集上固定公式 (2) 中的参数 λ_{uv} 时，参数 λ_u 和 λ_v 对 TRCF 性能的影响。我们可以看到，当 $\lambda_u = \lambda_v = 10$ 时，TRCF 达到最佳性能，这说明用户和物品语义信息都对模型性能有显着的贡献。图 4.4(b) 显示了在四个 DS-WO-CRI 子数据集上固定 λ_u 和 λ_v 为最佳值时，公式 (2) 中的参数 λ_{uv} 对 TRCF 性能的影响。我们可以看到，TRCF 的性能首先随着 λ_{uv} 的增大而上升，然后在某个阈值后开始下降。在 LC1、LC2、LC3、LC4 上最佳的 λ_{uv} 值分别是 0.0001、0.01、0.01 和 0.1。这个结果说明了 DS-WO-CRI 的程度越大，对应的 λ_{uv} 的最优值相应的越大。换句话来说，当 DS-WO-CRI 的问题越严重时，通过标签和评分（即隐式偏好）桥接用户和物品的特性越重要，这就解释了为什么我们的模型在可以在严重的 DS-WO-CRI 的情况下表现良好的原因。

5　总结

在本文中，我们首先介绍了在实际的标记系统中存在的 DS-WO-CRI 问题。然后，我们提出一个新型的基于标签和评分的 CF 模型来处理这个问题。该模型使用主题建模来分别挖掘用户和物品的标签的语义信息，并将语义信息结合到矩阵分解中以对评分信息进行因式分解，并捕获用户与物品之间的标签和评分的桥接特征。



(a) λ_u, λ_v 的影响　　(b) λ_{uv} 的影响

图 4.4　固定 M=50，$\lambda_u, \lambda_v, \lambda_{uv}$ 对 TRCF 召回率的影响。

据我们所知，这是文献中首次尝试引入 DS-WO-CRI 问题并提出一个处理它的模型。最后，对两个流行的数据集进行的实验表明，我们的模型在精确率和召回率方面显着优于目前最先进的方法，特别是在 DS-WO-CRI 情况下。

6　致谢

我们感谢匿名审稿人的有益建议。这项工作得到了中国国家自然科学基金（61379034）、国家重点技术研发计划（2014BAH28F05）、广东省科技计划（2013B040100004, 2013B040403002）和中国浙江省自然科学基金（LQ14F010006）的部分支持。

参 考 文 献

[1] Deepak Agarwal and Bee Chung Chen. flda:matrix factorization through latent dirichlet allocation. In International Conference on Web Search and Web Data Mining, WSDM 2010, New York, Ny, Usa, February, pages 91–100, 2010.

[2] Alejandro Bellogín, Iván Cantador, and Pablo Castells. A comparative study of heterogeneous item recommendations in social systems. Information Sciences An International Journal, 221(1):142–169, 2013.

[3] Dimitri P Bertsekas. Nonlinear programming. Athena scientific Belmont, 1999.

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

[5] Ivan Cantador, Peter Brusilovsky, and Tsvi Kuflik. Second workshop on information heterogeneity and fusion in recommender systems (hetrec2011). In ACM Conference on Recommender Systems, Recsys 2011, Chicago, Il, Usa, October, pages 387–388, 2011.

[6] C. Chen, X. Zheng, Y. Wang, F. Hong, and Z. Lin. Context-aware collaborative topic regression with social matrix factorization for recommender systems. Palo Alto California Aaai Press, 3(3):239–242, 2014.

[7] Chaochao Chen, Jing Zeng, Xiaolin Zheng, and Deren Chen. Recommender system based on social trust relationships. In IEEE International Conference on E-Business Engineering, pages 32–37, 2013.

[8] Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. ACM Transactions on Information Systems (TOIS), 22(1):143–177, 2004.

[9] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1):5–53, 2004.

[10] Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In ACM Conference on Recommender Systems, Recsys 2010, Barcelona, Spain, September, pages 135–142, 2010.

[11] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 426–434, 2008.

[12] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques

for recommender systems. ieee, computer journal, 42(8), 30-37. 42(8):30–37, 2009.

[13] Hao Ma, Dengyong Zhou, Chao Liu, Michael R. Lyu, and Irwin King. Recommender systems with social regularization. In Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February, pages 287–296, 2011.

[14] Prem Melville, Raymod J. Mooney, and Ramadass Nagarajan. Content-boosted collaborative filtering for improved recommendations. In Eighteenth National Conference on Artificial Intelligence, pages 187–192, 2002.

[15] Sanjay Purushotham, Yan Liu, and C-C Jay Kuo. Collaborative topic regression with social matrix factorization for recommendation systems. arXiv preprint arXiv:1206.4684, 2012.

[16] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In International Conference on Machine Learning, pages 880–887, 2008.

[17] Yue Shi, Martha Larson, and Alan Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Computing Surveys (CSUR), 47(1):3, 2014.

[18] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August, pages 448–456, 2011.

[19] Hao Wang, Binyi Chen, and Wu Jun Li. Collaborative topic regression with social regularization for tag recommendation. In International Joint Conference on Artificial Intelligence, pages 2719–2725, 2013.

[20] Jingwei Xu, Yuan Yao, Hanghang Tong, Xianping Tao, and Jian Lu. Ice-breaking: mitigating cold-start recommendation problem by rating comparison. In International Conference on Artificial Intelligence, pages 3981–3987, 2015.

[21] Nan Zheng and Qiudan Li. A recommender system based on tag and time information for social tagging systems. Expert Systems with Applications, 38(4):4575–4587, 2011.

[22] T. C Zhou, Hao Ma, I King, and M. R Lyu. Tagrec: Leveraging tagging wisdom for recommendation. In International Conference on Computational Science and Engineering, pages 194–199, 2009.

[23] Tom Chao Zhou, Hao Ma, Michael R. Lyu, and Irwin King. Userrec: A user recommendation framework in social tagging systems. In Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, Usa, July, pages 1486–1491, 2010.

Capturing Semantic Correlation for Item Recommendation in Tagging Systems

Chaochao Chen, Xiaolin Zheng
College of Computer Science
Zhejiang University
{zjuccc,xlzheng}@zju.edu.cn

Yan Wang
Department of Computing
Macquarie University
yan.wang@mq.edu.au

Fuxing Hong, Deren Chen
College of Computer Science
Zhejiang University
{cstur4,drc}@zju.edu.cn

Abstract

The popularity of tagging systems provides a great opportunity to improve the performance of item recommendation. Although existing approaches use topic modeling to mine the semantic information of items by grouping the tags labelled for items, they overlook an important property that tags link users and items as a bridge. Thus these methods cannot deal with the *data sparsity without commonly rated items* (*DS-WO-CRI*) problem, limiting their recommendation performance. Towards solving this challenging problem, we propose a novel tag and rating based collaborative filtering (CF) model for item recommendation, which first uses topic modeling to mine the semantic information of tags for each user and for each item respectively, and then incorporates the semantic information into matrix factorization to factorize rating information and to capture the bridging feature of tags and ratings between users and items. As a result, our model captures the semantic correlation between users and items, and is able to greatly improve recommendation performance, especially in *DS-WO-CRI* situations. Experiments conducted on two popular real-world datasets demonstrate that our proposed model significantly outperforms the conventional CF approach, the state-of-the-art social relation based CF approach, and the state-of-the-art topic modeling based CF approaches in terms of both precision and recall, and it is an effective approach to the *DS-WO-CRI* problem.

Introduction

In recent years, tagging systems, such as *Delicious* (social bookmarking), *Last.fm* (social music), *Flickr* (photo sharing), and *YouTube* (video sharing), provide effective ways for users to organize, manage, share, and search various kinds of items (resources). For example, one may tag Lady Gaga with "pop" and "female vocalist" when he listens to her music on *Last.fm*. These valuable tags, which appear along with the tagging and rating behaviors, strongly suggest the need to use such information to provide personalized recommendation services (Zheng and Li, 2011).

The increasing popularity of tagging systems has promoted the development of recommender systems, especially collaborative filtering (CF) approaches, in tagging systems. So far, two main types of CF on tagging systems exist: *tag recommendation* (Wang, Chen, and Li, 2013; Fang et al., 2015), which aims to recommend appropriate tags for items, and *tag-based item recommendation* (Zhou et al., 2009; Xu et al., 2011; Zhou et al., 2010), which focuses on recommending similar users or items to the target user based on tag and other information (e.g., rating).

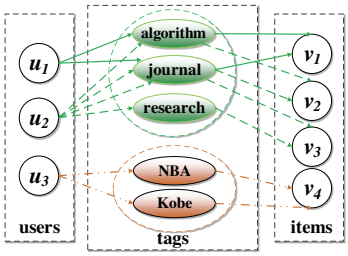


Figure 1: An example of tagging system. There is a rating behind each tagging behavior, and we omit the ratings for conciseness.

Currently, a trend in the literature is the use of topic modeling in CF to handle tag information (Agarwal and Chen, 2010; Wang and Blei, 2011; Purushotham, Liu, and Kuo, 2012; Wang, Chen, and Li, 2013; Chen et al., 2014). For example, Wang and Blei (2011) proposed a collaborative topic regression (CTR) model that can be used for tag-based item recommendation. Chen et al. (2014) proposed another item recommendation method that combined CTR with social matrix factorization (Ma et al., 2011) to make a better prediction. However, the existing approaches just associate tags with users or items, and overlook an important property of tags that tags link users and items as a bridge, as what ratings do. But tags can reflect the semantic correlation between users and items, which ratings cannot do.

When a user has tagged some items, these tags clearly represent the user's preference for the items. The more frequently a tag has been used by a user, the more likely this user is interested in the group of items that can be labeled by this tag (Zheng and Li, 2011). Similarly, the more frequently a tag has been given to an item by users, the more likely this item matches the tag. Thus, tags contain the semantic information of both users and items, not just one of them.

A motivating example. Figure 1 depicts an example of a

tagging system, which consists of three users (u_1, u_2, and u_3), four items (v_1, v_2, v_3, and v_4), and five tags ("algorithm", "journal", "research", "NBA", and "Kobe").

In this example, user u_1 labelled item v_1, and user u_2 labelled items v_2 and v_3. Thus users u_1 and u_2 have no commonly rated items. We term this situation *data sparsity without commonly rated items* (*DS-WO-CRI*). DS-WO-CRI is a typical subset of the standard data sparsity problem (i.e., the known user-item actions are rare comparing with all the user-item pairs). In *DS-WO-CRI* situations, the existing CF approaches, e.g., PMF and CTR, cannot essentially recommend item v_1 to user u_2 because they cannot capture any relation between them. However, a good recommender system should recommend items v_2 and v_3 to user u_1 and recommend item v_1 to user u_2, because in this example, users u_1 and u_2 are probably researchers on algorithms, and items v_1, v_2, and v_3 are probably related to algorithms.

The existing studies have shown that a user's action on an item, e.g., tag and rate, has already indicated this user's interests in this item, regardless of how the user rated this item (Koren, 2008; Koren, Bell, and Volinsky, 2009). In other words, a user implicitly expresses his preferences by voicing his opinion through tagging and voting a (high or low) rating (Koren, 2008). Thus, a user and the items that he has tagged and rated tend to share similar latent features, and we term it *implicit preference* in this paper. In the above example, in particular, user u_2 gives item v_3 tags "journal" and "research" and this not only shows that user u_2 is likely to be a researcher, but also indicates that item v_3 is likely to be a research journal or something related to it. Thus, semantically, the latent features of user u_2 and item v_3 should be similar to some extent.

However, the existing approaches fail to capture the semantic correlation between users and items, and thus their recommendation performance is limited, especially in *DS-WO-CRI* situations.

Our proposal. To deal with the above mentioned *DS-WO-CRI* problem, in this paper, we propose a novel CF model. We first use topic modeling to mine the semantic information of tags for each user and for each item respectively, and then incorporate the semantic information into matrix factorization to factorize rating information and capture the bridging feature of tags and ratings between users and items (i.e., implicit preference). As a result, our model captures the semantic correlation between users and items, and can recommend an item to a user if they have similar semantic information, though they are in a *DS-WO-CRI* situation.

Contributions. The main contributions of our work are summarized as follows: (1) We first point out the important feature of tags, namely, they link users and items as a bridge, outlining the semantic correlation between users and items, and then we illustrate that utilizing this feature can help deal with the *DS-WO-CRI* problem. To the best of our knowledge, this is the first study in the literature to identify this problem; (2) We propose a novel tag and rating based CF model, which can capture the semantic correlation between users and items and thus can greatly improve the recommendation performance, especially in *DS-WO-CRI* situations. We also propose our parameter learning method based on coordinate ascent algorithm. To the best of our knowledge, this study is the first attempt in the literature to capture the semantic correlation between users and items provided by tags in tag-based item recommendation; (3) Experiments conducted on two popular real-world datasets demonstrate that our proposed model significantly outperforms the state-of-the-art approaches in terms of both precision and recall. The experiments also demonstrate that our proposed model is an effective approach to the *DS-WO-CRI* problem.

Related Work

In this section, we review the existing item recommendation methods in tagging systems in three groups, including (1) the conventional CF approaches, (2) the topic modeling based CF approaches, and (3) the social relation based CF approaches.

Based on the existing research (Shi, Larson, and Hanjalic, 2014), the conventional CF approaches, which only use user-item rating information to make recommendations, are in two major categories: the memory-based CF (Deshpande and Karypis, 2004) and model-based CF (Koren, Bell, and Volinsky, 2009; Koren, 2008; Zhou et al., 2009; Xu et al., 2015), both of which can be used to make recommendations in tagging systems.

The conventional CF approaches, e.g., TagRec (Zhou et al., 2009), cannot capture content (e.g., tag) information. Thus, some hybrid approaches were proposed to combine content-based approach and CF to do item recommendation (Melville, Mooney, and Nagarajan, 2002) and (Basilico and Hofmann, 2004). However, these methods take content simply as a vector of words, and thus cannot mine their semantic information. To take advantage of semantic information provided by content (e.g., tag), researchers use topic modeling to improve recommendation performance. Agarwal et al., proposed fLDA (Agarwal and Chen, 2010), which combines RLFM (Agarwal and Chen, 2009) with latent Dirichlet allocation (LDA) by assigning item factors through a richer prior learnt from LDA. Both RLFM and fLDA incorporate additional covariates that are obtained from additional meta-feature information, e.g., user age and item category, which, however, are out of the scope of this paper. Later on, Wang et al. (Wang and Blei, 2011) proposed CTR to combine probabilistic matrix factorization (PMF) (Mnih and Salakhutdinov, 2007) with LDA (Blei, Ng, and Jordan, 2003) to make recommendations. It has been proven in (Wang and Blei, 2011) that CTR performs better than fLDA in a similar setting, since fLDA largely ignores the other users' ratings.

Moreover, social information between users and between items is considered valuable to improve recommendation performance (Chen et al., 2013). First, user social information is incorporated into conventional CF models (Jamali and Ester, 2010; Ma et al., 2011). For example, Ma et al. (2011) proposed Soreg to constrain the difference between the user latent factors of connected users. Second, neighbor user or item social information is incorporated into topic modeling based CF models (e.g., CTR) to further improve recommendation performance. For example, Purushotham, Liu, and Kuo (2012) and Chen et al. (2014) proposed two

models (i.e., CTR-SMF and CTR-SMF2) to incorporate user social network into CTR to further improve item recommendation performance. Wang, Chen, and Li (2013) proposed a model to incorporate item social relationship into CTR to further improve tag recommendation performance in social tagging systems.

However, all the above approaches not only overlook the semantic information between both users and items embedded in tags, but also neglect the bridging feature of tags and ratings between users and items (i.e., implicit preference). Therefore, they cannot capture the semantic correlation between users and items, and suffer from *DS-WO-CRI* problem. To overcome these shortcomings, in this paper, we propose a novel tag and rating based CF model, which can capture the semantic correlation between users and items. Hence, our model can help deal with the *DS-WO-CRI* problem and improve recommendation performance.

The Proposed Model–TRCF

In this section, we present a novel tag and rating based CF (TRCF) model. We first formalize the tag-based item recommendation problem and define notations. Then, we present TRCF, which is a hierarchical Bayesian model. Finally, we propose our parameter learning method based on coordinate ascent algorithm.

Preliminaries

Assume that we have a set of users $\mathbb{U} = \{u_1, ..., u_I\}$, who have labelled a set of items $\mathbb{V} = \{v_1, ..., v_J\}$ with a set of tags $\mathbb{T} = \{t_1, ..., t_N\}$ and a set of ratings $\mathbb{R} = \{R_1, ..., R_O\}$, where *I*, *J*, *N*, and *O* denote the numbers of users, items, tags, and ratings, respectively. Each *user-item-tag-rating* (U-I-T-R) observe data is a 4-tuple (u_i, v_j, T_{ij}, R_{ij}), where $u_i \in \mathbb{U}$, $v_j \in \mathbb{V}$, T_{ij} is a set of tags that user u_i gives to item v_j, and $T_{ij} \subseteq \mathbb{T}$. R_{ij} is the rating that user u_i gives to item v_j based on the extent to which he likes the item and tags it at the same time; however, the user-item (U-I) rating set $\mathbb{R}$ is typically of integers, e.g., in the range [1, 5] in *MovieLens*. Let $U \in \mathbb{R}^{K \times I}$ denote the latent user feature matrices, where the column vector U_i represents the *K*-dimensional user-specific latent feature vector of user u_i. Let $V \in \mathbb{R}^{K \times J}$ denote the latent item feature matrices, where the column vector V_j represents the *K*-dimensional item-specific latent feature vector of item v_j.

For tag and rating based item recommendation, given the existing U-I-T-R 4-tuples, our goal is to predict the unknown rating from a user u_i to an item v_j.

Tag and Rating based Collaborative Filtering

TRCF is a novel hierarchical Bayesian model, and its graphical model is shown in Figure 2, where N_u and N_v denote the number of tags for user u_i and for item v_j, respectively. TRCF first groups the tags for each user and for each item respectively, and then it uses latent Dirichlet allocation (LDA) to mine the semantic information of tags for each user and each item respectively (plotted in red in Figure 2). Finally it incorporates these semantic information into matrix factorization to factorize rating information (plotted in purple in Figure 2) and capture the implicit preference provided by tags and ratings (plotted in blue in Figure 2).

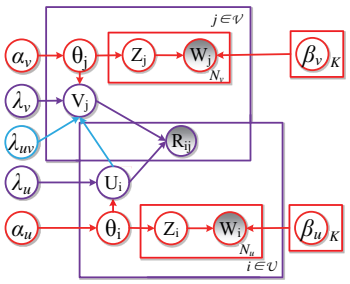


Figure 2: Graphical model of TRCF. The LDA part is plotted in red, the implicit preference part is shown in blue, and the PMF part is plotted in purple.

TRCF performs LDA on both the user side and the item side, and thus can capture the semantic information for both users and items, not just items as the existing works do. In addition, in TRCF, when a user and an item are linked by tags and ratings, their latent features are similar to each other to some extent, which is referred to as implicit preference. In contrast, the existing topic modeling based CF approaches, e.g., CTR, CTR-SFM, and CTR-SMF2, assume users and items are independent, and neglect the bridge feature of tags and ratings between users and items. Thus, TRCF can capture the semantic correlation between users and items, and is able to deal with the *DS-WO-CRI* problem. Assuming there are *K* topics for both users and items, the generative process of TRCF works as follows:

1. **Mining semantic information of tags for users**. For each user u_i,
   (a) Draw topic proportions $\theta_i \sim Dirichlet(\alpha_u)$;
   (b) Draw user latent vector as $U_i \sim \mathcal{N}(\theta_i, \lambda_u^{-1} I_K)$;
   (c) For each tag w_{in_u} of user u_i,
      i. Draw topic assignment $z_{in_u} \sim Mult(\theta_i)$;
      ii. Draw tag $w_{in_u} \sim Mult(\beta_{z_{in_u}})$;
2. **Mining semantic information of tags for items, and capturing implicit preference between users and items**. For each item v_j,
   (a) Draw topic proportions $\theta_j \sim Dirichlet(\alpha_v)$;
   (b) Draw item latent vector as $V_j \sim \mathcal{N}(\theta_j, \lambda_v^{-1} I_K) \times \prod_i I_{ij}^R \mathcal{N}(U_i, \lambda_{uv}^{-1} I_K)$;
   (c) For each tag w_{jn_v} of user v_j,
      i. Draw topic assignment $z_{jn_v} \sim Mult(\theta_j)$;
      ii. Draw tag $w_{jn_v} \sim Mult(\beta_{z_{jn_v}})$;
3. **Drawing the rating**. For each user-item pair (*i*, *j*),

$$R_{ij} \sim \mathcal{N}({U_i}^T V_j, c_{ij}^{-1}).$$

In the above generative process, $\mathcal{N}(x|\mu, \sigma^2)$ is a Gaussian distribution with a mean μ and a variance σ^2, and I_K is an identity matrix with K rows and K columns. I_{ij}^R is an indicator function the value of which equal to 1 if user u_i

rated item v_j, 0 otherwise. C is a rating confidence matrix with element c_{ij} denotes the rating confidence. Please refer to (Wang and Blei, 2011) for more details.

The parameter λ_u balances the contribution of user semantic information provided tags and rating information to the model performance. Similarly, the parameter λ_v balances the contribution of item semantic information provided by tags and rating information to the recommendation performance. The parameter λ_{uv} balances the contribution of implicit preference on model performance, i.e., the degree of the latent feature similarity of a user and an item linked by a rating and tags.

The conditional distribution of the observed ratings can be formalized as

$$p\left(R|U,V,C\right)=\prod_i\prod_j\mathcal{N}(R_{ij}|U_i^TV_j,c_{ij}).$$

The user and latent vectors U_i and V_j are generated in a similar way to CTR, which can be formalized as

$$p\left(U|\lambda_u\right)\sim\prod_i\mathcal{N}(\theta_i,\lambda_u^{-1}I_K),$$

$$p\left(V|U,\lambda_v,\lambda_{uv}\right)\sim\prod_j\mathcal{N}(\theta_j,\lambda_v^{-1}I_K)\times\prod_i I_{ij}^R\mathcal{N}(U_i,\lambda_{uv}^{-1}I_K).$$

Given the U-I-T-R information, by using Bayesian inference, we can obtain the following equation for the posterior probability of latent feature vectors of TRCF:

$$\begin{aligned}&p\left(U,V|R,C,\lambda_u,\lambda_v,\lambda_{uv}\right)\\&\propto p\left(R|U,V,C\right)p\left(U|\lambda_u\right)p\left(V|U,\lambda_v,\lambda_{uv}\right).\end{aligned}\tag{1}$$

Parameter Learning of TRCF

Given topic parameters β_u and β_v, computing the full posterior of U_i, V_j, θ_i, and θ_j directly is intractable. We use coordinate ascent algorithm to learn the maximum a posteriori estimates. Maximizing the posterior over the two latent features with fixed hyper-parameters in Equation (1) is equivalent to maximizing the following complete log likelihood of U, V, $\theta_{1:I}$, $\theta_{1:J}$ and R, given λ_u and λ_v:

$$\begin{aligned}L=&-\frac{\lambda_u}{2}\sum_i\left(U_i-\theta_i\right)^T\left(U_i-\theta_i\right)\\&-\frac{\lambda_v}{2}\sum_j\left(V_j-\theta_j\right)^T\left(V_j-\theta_j\right)\\&-\sum_{ij}\frac{c_{ij}}{2}\left(R_{ij}-{U_i}^TV_j\right)^2\\&+\sum_i\sum_{n_u}\log\left(\sum_k\theta_{ik}\beta_{k,w_{in_u}}\right)\\&+\sum_j\sum_{n_v}\log\left(\sum_k\theta_{jk}\beta_{k,w_{jn_v}}\right)\\&-\frac{\lambda_{uv}}{2}I_{ij}^R\sum_{ij}(U_i-V_j)^T(U_i-V_j).\end{aligned}\tag{2}$$

We omit a constant and set the Dirichlet priors $\alpha_u=\alpha_v=1$. This function can be optimized by using coordinate ascent. That is, we fix β_u and β_v, and iteratively optimize the MF variables U_i, V_j and the topic proportions θ_i and θ_j. Specifically, we first update U_i and V_j, given the current estimate of θ_i, θ_j. We take the gradient of L in Equation (2) with respect to U_i and V_j, and set it to zero,

$$\frac{\partial L}{\partial U_i}=0,\frac{\partial L}{\partial V_j}=0.\tag{3}$$

Solving the above equations will lead to the following update equation,

$$\begin{aligned}U_i\leftarrow&\left(VC_iV^T+\lambda_uI_K+\lambda_{uv}\sum_jI_{ij}^RI_K\right)^{-1}\\&(VC_iR_i+\lambda_u\theta_i+\lambda_{uv}\sum_jI_{ij}^RV_j),\\V_j\leftarrow&\left(UC_jU^T+\lambda_vI_K+\lambda_{uv}\sum_iI_{ij}^RI_K\right)^{-1}\\&(UC_jR_j+\lambda_v\theta_j+\lambda_{uv}\sum_iI_{ij}^RU_i),\end{aligned}\tag{4}$$

where C_i is a diagonal matrix with $c_{ij}, j=1,...,J$ as its diagonal elements, and $R_i=R_{ij}{}_{j=1}^{J}$ for user u_i. For item v_j, C_j and R_j are similarly defined.

Equation (4) shows how parameters λ_u, λ_v, and λ_{uv} affect the user latent feature and the item latent feature. A bigger λ_u corresponds to a bigger proportion of the user latent feature from the user tags rather than the rating information. Similarly, a bigger λ_v indicates a bigger proportion of the item latent feature from the item tags, rather than the rating information. Also, a bigger λ_{uv} means a stronger constraint that the paired user and item linked by tags and ratings should have a similar latent feature, i.e., implicit preference. From Equation (4), we can see that probabilistic matrix factorization (PMF) and collaborative topic regression (CTR) are all special cases of TRCF.

Then, we update the topic proportions θ_i and θ_j given the current MF variables U_i and V_j. For θ_i, we first define $q\left(z_{in_u}=k\right)=\Phi_{in_uk}$, and then separate the users that contain θ_i and apply Jensen's inequality,

$$\begin{aligned}L(\theta_i)\geq&-\frac{\lambda_u}{2}(U_i-\theta_i)^T(U_i-\theta_i)\\&+\sum_{n_u}\sum_k\Phi_{in_uk}(log\theta_{ik}\beta_{k,w_{in_u}}-log\Phi_{in_uk})\\=&L(\theta_i,\Phi_i).\end{aligned}$$

Here, $\Phi_i=\Phi_{in_uk}{}_{n_u=1,k=1}^{N_u\times K}$. Obviously $L(\theta_i,\Phi_i)$ is a tight lower bound of $L(\theta_i)$, and we can use projection gradient (Bertsekas, 1995) to optimize θ_i. The optimal Φ_{in_uk} is $\Phi_{in_uk}\propto\theta_{ik}\beta_{k,w_{in_u}}$. For θ_j, it is similarly updated.

As for β_u, we update the same M-step for topics as in LDA (Blei, Ng, and Jordan, 2003),

$$\beta_{kw_i}\propto\sum_i\sum_{n_u}\Phi_{in_uk}1[w_{in_u}=w].$$

For β_v [1], it is similarly updated. After the optimal param-

[1] When using TF-CTR, a useful tactic is to fuse the user tags and item tags as the input of LDA, which ensures that users and items have the same semantic information in each element of K, that is, to make $\beta_u=\beta_v$.

latex表格

Dataset	users	items	tags	user-tags-items	ratings
Delicious	1,867	69,226	53,388	437,593	104,799
Lastfm	1,892	17,632	11,946	186,479	92,834

Table 1: Dataset description

eters U^*, V^*, ${\theta_{1:I}}^*$, ${\theta_{1:J}}^*$, ${\beta_u}^*$, and ${\beta_v}^*$ have been learned, our model can predict ratings:

$$ {R_{ij}}^* \approx (U_i^*)^T V_j^*. $$

Experiments and Analysis

In this section, we introduce the experiments conducted on two popular real-world datasets, which aim to answer the following questions: (1) How does our model perform comparing the state-of-the-art approaches? (2) How does our approach deal with the *DS-WO-CRI* problem? (3) How do parameters λ_u, λ_v, and λ_{uv} affect the performance of TRCF?

Datasets

We use two real-world datasets in our experiments: hetrec2011-delicious-2k (*Delicious*) and hetrec2011-lastfm-2k (*Lastfm*) (Cantador, Brusilovsky, and Kuflik, 2011). Both datasets have been widely used to conduct experiments in tagging systems (Bellogín, Cantador, and Castells, 2013), and they are described in Table 1.

For each of the two datasets, we consider a user rating for an item as '1' if the user has bookmarked (or listened) the item; otherwise, the user rating for the item is '0'.

In our experiments, we split each dataset into three parts — a training dataset (80%), a held-out validation dataset (10%), and a test dataset (10%). We train our model on the training dataset, obtain the optimal parameters on the validation dataset, and evaluate our model on the test dataset.

Comparison and Evaluation

As stated in related works, there are many kinds of recommendation approaches, e.g., memory-based approach and hybrid approach. Here, we compare the proposed TRCF with the following three kinds of state-of-the-art approaches, i.e., the conventional CF approach, the social relation based CF approach, and the topic modeling based approach:

SVD++ (Koren, 2008) is a classic conventional CF approach that only uses U-I rating information.

Soreg (Ma et al., 2011) is a state-of-the-art social relation based CF approach, which uses U-I rating and user social information.

CTR (Wang and Blei, 2011) is a state-of-the-art topic modeling based CF approach, which uses U-I-T-R information similar to the 4-tuple used in TRCF.

CTR-SMF (Purushotham, Liu, and Kuo, 2012) combines user social matrix factorization with CTR. It incorporates additional user social information additional to the U-I-T-R 4-tuple used in TRCF.

CTR-SMF2 (Chen et al., 2014) improves CTR-SMF, and it also incorporates user social information additional to the U-I-T-R 4-tuple used in TRCF.

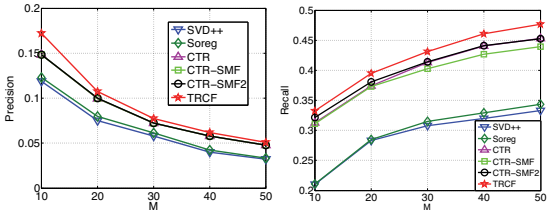


Figure 3: *Precision* and *Recall* comparison with different *M* and the best parameters of each method on *Delicious*.

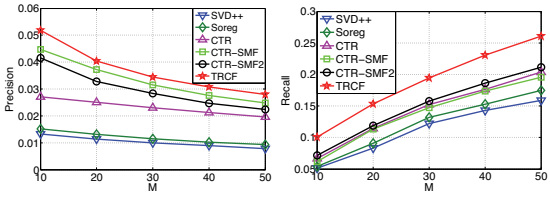


Figure 4: *Precision* and *Recall* comparison with different *M* and the best parameters of each method on *Lastfm*.

Precision and *Recall* have been widely used as the metrics to evaluate recommendation performance (Herlocker et al., 2004). Thus, we use both *Precision* and *Recall* to evaluate the recommendation performance, and note that the same way of computing recall is also used in CTR, CTR-SMF, and CTR-SMF2. For each user, *Precision* and *Recall* are defined as follows:

$$ Precision@M = \frac{\#\ items\ the\ user\ likes\ in\ Top\ M}{M}, $$

$$ Recall@M = \frac{\#\ items\ the\ user\ likes\ in\ Top\ M}{\#\ total\ items\ the\ user\ likes}, $$

where *M* is the number of returned items. We compute the average of all the items' precision and recall in the test dataset as the final result.

Performance Comparison and Analysis

During the comparison, we have used the best parameters for SVD++, CTR, and CTR-SMF that are set in CTR-SMF (Purushotham, Liu, and Kuo, 2012), which uses the same datasets. For each of Soreg, CTR-SMF2, and our model, we have used grid search to obtain the best parameters.

Results: Figures 3 and 4 show the overall performance for each recommendation approach on *Delicious* dataset and *Lastfm* dataset, in which we set *M* = 10, 20, 30, 40, 50 and fix the parameters of each approach to the best values. The results show that the conventional CF approach (i.e., SVD++) and the social relation based CF approach (i.e., Soreg) have the similar performance. The three topic modeling based CF approaches (i.e., CTR, CTR-SMF, and CTR-SMF2) significantly outperform SVD++ and Soreg, and also have similar performance, which shows the importance of tag information in recommendations.

Our proposed method, TRCF, significantly outperforms each of SVD++, Soreg, CTR, CTR-SMF, and CTR-SMF2

sub-dataset name	*LC1*	*LC2*	*LC3*	*LC4*
DS-WO-CRI degree	20%	40%	60%	80%
users	1,837	1,709	1,718	1,706
items	11,584	8,018	7,650	7,431

Table 2: Statistics of each *DS-WO-CRI* sub-dataset

on the two datasets in terms of different *M*. Specifically, on average, TRCF improves SVD++, Soreg, CTR, CTR-SMF, and CTR-SMF2 by 46.75%, 39.74%, 8.62%, 8.78%, and 8.65%, in terms of precision, and by 44.99%, 42.40%, 5.23%, 7.27%, and 4.21%, in terms of recall, on the *Delicious* dataset. On average, TRCF improves SVD++, Soreg, CTR, CTR-SMF, and CTR-SMF2 by 259.80%, 210.27%, 57.96%, 11.64%, and 23.85%, in terms of precision, and by 73.03%, 60.39%, 34.18%, 39.04%, and 28.12%, in terms of recall, on the *Lastfm* dataset.

Analysis and summary: The comparison demonstrates the effectiveness of our proposed method which captures the semantic correlation between users and items. Experimental results also indicate that though user social information (e.g., adopted in Soreg, CTR-SMF, and CTR-SMF2) can improve recommendation performance, considering user and item semantic correlation is a more effective way to improve item recommendation performance.

DS-WO-CRI Experiments

All four topic modeling based CF approaches, including TRCF, can improve recommendation performance by capturing the semantic information of items. To study their capability of handing the *DS-WO-CRI* problem, we conduct the following experiments.

We first randomly filter the original *Lastfm* datasets into four sub-datasets based on the degree of *DS-WO-CRI*, and each sub-dataset is described in Table . The degree of *DS-WO-CRI* is defined as follows:

$$x\% = \frac{\#\ users\ without\ commonly\ rated\ items}{\#\ total\ users}.$$

We then conduct comparison experiments on each of the sub-dataset. Figure 5 shows that our model always achieves the best performance under different *DS-WO-CRI* degrees. The average improvements of our model over other three topic modeling based approaches on *LC1*, *LC2*, *LC3*, *LC4* are 49.67%, 75.81%, 345.77%, and 428.97% respectively, in terms of precision, and are 33.94%, 65.00%, 383.63%, and 458.00% respectively, in term of recall. The experimental results show that a greater degree of *DS-WO-CRI* corresponds to a higher improvement of our model against other models on each of the two datasets.

Summary: The *DS-WO-CRI* experiments demonstrate the effectiveness of our method in dealing with the *DS-WO-CRI* problem: a greater degree of *DS-WO-CRI* corresponds to a higher improvement of our model against other models. This is due to the ability of our method to capture the semantic correlation between users and items provided by tags.

Parameter Effect Analysis

Figure 6(a) shows the effect of λ_u and λ_v in Eq.(2) on the performance of TRCF by fixing $\lambda_{uv} = 0$ on *Lastfm*. We can see that TRCF achieves the best performance when $\lambda_u = \lambda_v = 10$, which means that both user and item semantic information contribute significantly to model performance. Figure 6(b) shows how the performance of TRCF is affected by parameter λ_{uv} in Eq.(2) on each *DS-WO-CRI* sub-dataset with the best λ_u and λ_v. As we can see, the performance of TRCF first increases with λ_{uv} and then starts to decrease at a certain threshold. The best value of λ_{uv} on *LC1*, *LC2*, *LC3*, and *LC4* is 0.0001, 0.01, 0,01, and 0.1, respectively. These results demonstrate that a greater degree of *DS-WO-CRI* corresponds to a greater λ_{uv}. In other words, the feature of bridging users and items by tags and ratings (i.e., the implicit preference) is more important when the *DS-WO-CRI* problem is severer, which explains why our model performs well in severe *DS-WO-CRI* situations.

Conclusions

In this paper, we first introduce the *DS-WO-CRI* problem that exists in item recommendation in real tagging systems. Then, we present a novel tag and rating based CF model to deal with this problem. The proposed model uses topic modeling to mine the semantic information of tags for users and items respectively, and incorporates the semantic information into matrix factorization to factorize the rating information and to capture the bridging feature of tags and ratings between users and items. To the best of our knowledge, this is the first attempt in the literature to introduce the *DS-WO-CRI* problem and propose a model to deal with it. Finally, the experiments conducted on two well-known datasets have demonstrated that our model significantly outperforms the state-of-the-art approaches in terms of both precision and recall, especially in *DS-WO-CRI* situations.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. This work was supported in part by the National Natural Science Foundation of China (No. 61379034), the National Key Technology R&D Program (No. 2014BAH28F05), the Guangdong Province Science and Technology Program (No. 2013B040100004 and No. 2013B040403002), and the Zhejiang Provincial Natural Science Foundation of China (LQ14F010006).

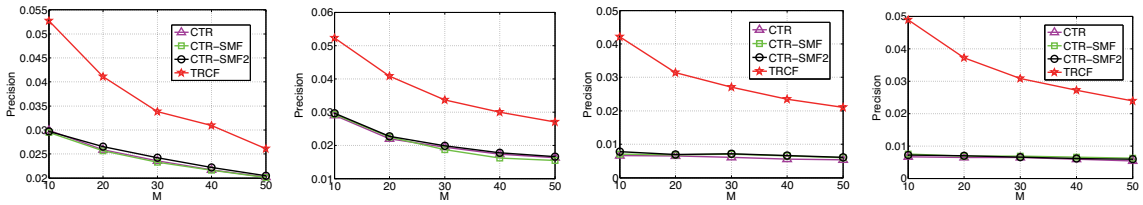
References

Agarwal, D., and Chen, B.-C. 2009. Regression-based latent factor models. In *SIGKDD*, 19–28.

Agarwal, D., and Chen, B.-C. 2010. flda: matrix factorization through latent dirichlet allocation. In *WSDM*, 91–100.

Basilico, J., and Hofmann, T. 2004. Unifying collaborative and content-based filtering. In *ICML*, 65–72.

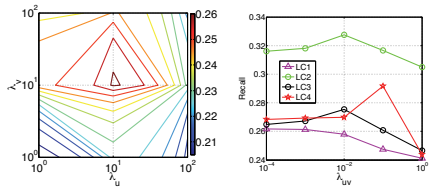
Bellogín, A.; Cantador, I.; and Castells, P. 2013. A comparative study of heterogeneous item recommendations in social systems. *Information Sciences* 221:142–169.



(a) *LC1* (20% *DS-WO-CRI*) (b) *LC2* (40% *DS-WO-CRI*) (c) *LC3* (60% *DS-WO-CRI*) (d) *LC4* (80% *DS-WO-CRI*)

(e) *LC1* (20% *DS-WO-CRI*) (f) *LC2* (40% *DS-WO-CRI*) (g) *LC3* (60% *DS-WO-CRI*) (h) *LC4* (80% *DS-WO-CRI*)

Figure 5: *Precision* and *Recall* comparison on each *DS-WO-CRI* sub-dataset with the best parameters of each method.



(a) Effect of λ_u and λ_v (b) Effect of λ_{uv}

Figure 6: Effect of λ_u, λ_v, and λ_{uv} on the recall performance of TRCF by fixing M=50.

Bertsekas, D. P. 1995. Nonlinear programming. *Athena Scientific*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR* 3:993–1022.

Cantador, I.; Brusilovsky, P.; and Kuflik, T. 2011. Second workshop on information heterogeneity and fusion in recommender systems (hetrec2011). In *RecSys*, 387–388.

Chen, C.; Zeng, J.; Zheng, X.; and Chen, D. 2013. Recommender system based on social trust relationships. In *ICEBE*, 32–37.

Chen, C.; Zheng, X.; Wang, Y.; Hong, F.; and Lin, Z. 2014. Context-aware collaborative topic regression with social matrix factorization for recommender systems. In *AAAI*. 9–15.

Deshpande, M., and Karypis, G. 2004. Item-based top-n recommendation algorithms. *TOIS* 22(1):143–177.

Fang, X.; Pan, R.; Cao, G.; He, X.; and Dai, W. 2015. Personalized tag recommendation through nonlinear tensor factorization using gaussian kernel. In *AAAI*, 439–445.

Herlocker, J. L.; Konstan, J. A.; Terveen, L. G.; and Riedl, J. T. 2004. Evaluating collaborative filtering recommender systems. *TOIS* 22(1):5–53.

Jamali, M., and Ester, M. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *RecSys*, 135–142.

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

Koren, Y. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *SIGKDD*, 426–434.

Ma, H.; Zhou, D.; Liu, C.; Lyu, M. R.; and King, I. 2011. Recommender systems with social regularization. In *WSDM*, 287–296.

Melville, P.; Mooney, R. J.; and Nagarajan, R. 2002. Content-boosted collaborative filtering for improved recommendations. In *AAAI*, 187–192.

Mnih, A., and Salakhutdinov, R. 2007. Probabilistic matrix factorization. In *NIPS*, 1257–1264.

Purushotham, S.; Liu, Y.; and Kuo, C.-c. J. 2012. Collaborative topic regression with social matrix factorization for recommendation systems. In *ICML*, 759–766.

Shi, Y.; Larson, M.; and Hanjalic, A. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *CSUR* 47(1):3.

Wang, C., and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *SIGKDD*, 448–456.

Wang, H.; Chen, B.; and Li, W.-J. 2013. Collaborative topic regression with social regularization for tag recommendation. In *IJCAI*, 2719–2725.

Xu, G.; Gu, Y.; Dolog, P.; Zhang, Y.; and Kitsuregawa, M. 2011. Semrec: a semantic enhancement framework for tag based recommendation. In *AAAI*. 1267–1272.

Xu, J.; Yao, Y.; Tong, H.; Tao, X.; and Lu, J. 2015. Ice-breaking: Mitigating cold-start recommendation problem by rating comparison. In *IJCAI*, 3981–3987.

Zheng, N., and Li, Q. 2011. A recommender system based on tag and time information for social tagging systems. *ESWA* 38(4):4575–4587.

Zhou, T. C.; Ma, H.; King, I.; and Lyu, M. R. 2009. Tagrec: Leveraging tagging wisdom for recommendation. In *CSE*, volume 4, 194–199.

Zhou, T. C.; Ma, H.; Lyu, M. R.; and King, I. 2010. Userrec: A user recommendation framework in social tagging systems. In *AAAI*. 1486–1491.

本科毕业论文文献综述

摘要 推荐系统通过分析用户的兴趣模式，在互联网海量数据中为用户提供个性化的产品和内容建议，大大降低了用户检索信息的成本。推荐系统是一个热门的研究领域，越来越多的与推荐系统相关的工作被发表。良好的个性化推荐对提升用户体验有很大帮助，近年来推荐系统被成功地运用在了许多电子商务和内容提供的网站。本文将讨论近年来推荐系统领域的热门研究，并总结推荐系统中冷启动和稀疏性等关键的问题。

1 背景

随着信息技术和互联网的发展，人们逐渐从信息匮乏的时代走入了信息过载的时代，海量信息的复杂性和不均匀性使得信息获取变得困难而耗时，无论信息消费者还是信息生产者都遇到了很大的挑战。关于信息过载的问题，代表性的解决方案是分类目录和搜索引擎，分类目录只能覆盖少量内容而越来越不能满足用户的需求，搜索引擎可以让用户通过搜索关键词找到自己需要的信息，但是，搜索引擎需要用户主动对需求提供描述，当用户无法找到准确描述自己需求的关键词时，搜索引擎就无能为力了。

与搜索引擎不同，推荐系统不需要用户提供明确的需求，而是通过分析用户的过往兴趣模式，自动为用户过滤掉低相关的内容，呈现符合他们兴趣和需求的个性化建议。从某种意义上来说，推荐系统和搜索引擎是两个互补的技术，推荐系统满足了用户有明确目的时的主动查找需求，而推荐系统能够在没有明确目的时帮助用户发现感兴趣的内容 [1]。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、用户的历史兴趣记录以及用户的注册信息等。

越来越多的网站成功地引入了推荐系统，在电子商务领域，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，提高了用户的购买效率，同时也为商家盈利带来好处；对于诸如音乐、电影、视频这样的娱乐内容，为用户推荐可能感兴趣的内容可以为内容提供商吸引更多的用户流量；互联网广告的个性化定向投放能准确的找到潜在客户群，相比于随机投放提高了效率。

推荐系统依赖于不同类型的用户行为数据，最理想的是高质量的显式反馈行为，即用户对物品兴趣的明确输入，主要的方式就是评分或单纯的喜不喜欢。通常，显式反馈产生稀疏的偏好度矩阵，因为单个用户可能只评价了一小部分物品。和显式反

馈行为相对应的是隐式反馈行为,即那些不能明确反应用户喜好的行为,例如购买商品、浏览页面、评论或甚至鼠标移动。相比显式反馈,隐式反馈虽然不明确,但数据量更大,因此利用隐式反馈缓解数据的稀疏问题。为了简单起见,我们将各种类型的反馈统称为评分(rating)。

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,而对于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。

概括地说,推荐系统主要基于两种不同的策略或其组合:基于内容的过滤方法和协同过滤方法。

内容过滤:基于内容的过滤方法为每个用户或物品创建描述以表征其性质,例如,电影的描述可以包括其类型、导演、票房等方面,用户的描述可以是个人资料或从已评分物品中识别出的共同特征。这样就可以将用户的描述做为关键词,利用信息检索的方式匹配物品的描述。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,利用物品的描述即可进行推荐。当然,缺点是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。

协同过滤:另一种策略,不像内容过滤那样需要明确的描述信息,而是基于用户的行为分析用户的兴趣,这种方法被称为协同过滤(Collaborative Filtering)。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。协同过滤克服了基于内容的一些限制,它比内容过滤的技术更加精确,但是却无法解决系统新用户和物品的冷启动问题,同时,稀疏且不均匀的历史数据使得分析变得困难。

预测的准确度是度量一个推荐系统预测能力的指标,计算该指标需要离线的包括用户历史行为记录的数据集,并将数据集按时间或随机地分为训练集和测试集,然后通过在训练集上建立用户的兴趣模型来预测用户在测试集上的行为,再计算预测行为和在测试集上的真实行为的重合度作为预测准确度。推荐的任务可以分为评分

预测和 TopN 推荐两种:很多网站有让用户给物品打分的功能,评分预测就是预测用户给他未评分的物品的评分,通常用平均绝对误差(MAE)和均方根误差(RMSE)来评估预测准确度 [2]。网站在提供推荐服务时一般会给用户一个个性化的推荐列表,这种推荐叫做 TopN 推荐,这种方式推荐的准确度一般利用召回率(Recall)和精确率(Precision)来度量。覆盖率(coverage)描述一个推荐系统对冷门物品的发掘能力,定义为推荐系统能够推荐出来的物品占总物品集合的比例。覆盖率对于内容提供商来说是一个重要的指标,一个好的推荐系统不仅需要有比较高的用户满意度,也要有较高的覆盖率。

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。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 [3]。本文将介绍推荐系统目前的研究现状,从基本的协同过滤方法开始讨论,并且围绕冷启动和矩阵稀疏性两个最主要的挑战,对近些年来的热门研究成果进行综述。

2 协同过滤的研究现状

推荐系统利用数据分析技术生成用户物品的预测偏好度,为用户找到那些可能喜欢的物品,通常推荐方法都是利用用户信息、热门物品、历史行为这些数据进行预测。基于内容的过滤方法创建用户和物品的描述,以此来选择符合用户特征的物品。然而纯粹基于内容的推荐系统通常导致推荐过于局限化的问题,即无法推荐丰富多样的物品给用户,另一方面,用于生成特征描述的信息也并不容易获得。

协同过滤 (Collaborative Filtering) 这个术语最早由 David Goldberg 等人于 1992 年创造,用于描述一个实验性的邮件过滤系统 Tapestry[4]。在该系统中,每个用户可以为每个邮件编写注释并且与一组用户共享这些注释,然后,用户可以通过对这些注释进行查询来过滤这些电子邮件。尽管 Tapestry 使用户受益于其他用户的注释,但该系统仍需要用户编写复杂的查询,之后随着推荐系统的发展,出现了自动化生成推荐的技术,最早的自动推荐系统是 GroupLens,它识别相似用户的集合,并筛选该集合中的物品来获得对每个个体的建议。

协同过滤推荐是推荐系统中应用最早和最为成功的策略之一,核心是围绕用户物品的偏好度矩阵展开的,分析用户之间的关系和物品之间的相互依赖性,以获得新的用户和物品的关联。协同过滤克服了基于内容过滤的一些限制,它不使用内容信

息，而是使用系统中其他用户和项目的评分信息。此外，与基于内容的系统不同，协作过滤可以推荐多样类型的物品，只要其他用户已经对这些不同物品表现出兴趣。

随着互联网的不断发展，尤其是电子商务的出现，推荐算法随之快速发展，协同过滤在研究领域和实践中都去得了巨大成功，目前大部分推荐算法都是基于协同过滤算法改进而来。协同过滤的两个主要领域是基于邻域的方法和潜在因素模型。在基于邻域的协同过滤中，存储在系统中的用户项目评分直接用于预测新项目的评分，因此有被称为基于存储的协同过滤。基于模型的方法不直接利用存储的评分进行预测，而是使用这些评分来训练预测模型，模型的参数捕获了用户和项目的潜在特征，这些模型参数从训练数据中学习而来并用于随后的预测。

2.1 基于邻域的协同过滤

最常见的协同过滤是基于邻域的方法，该方法的基本思想借鉴了人们生活中选择物品的方式，如果身边的朋友喜欢某件物品，那么自己就会有很大概率选择该物品。另外，如果用户喜欢某个物品，那么他很可能喜欢与该物品类似的物品。因此，该方法又分为基于用户的方法和基于物品的方法。

基于邻域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。

2.1.1 基于用户的协同过滤

协同过滤最初的形式是以用户之间的关系为中心 [5]，系统的实现分为两个阶段，首先在历史数据中发现那些具有相似品味的用户，然后利用邻居对物品 i 的评分计算用户 u 对一个新物品 i 的评分 r_{ui} 。假设我们计算得到了每个用户对 $u \neq v$ 之间的相似度 w_{uv} ，与用户 u 相似度最高的 k 个用户的集合，称为 u 的 k 近邻（k -NN），该集合记为 $N(u)$。然而只有评价过物品 i 的邻居才可以用来预测 r_{ui} ，因此我们将这个邻居的集合记为 $N_i(u)$ ，可以用这些邻居对 i 评分的平均值来估计 r_{ui} ：

$$\hat{r_{ui}} = \frac{1}{|N_i(u)|} \sum_{v \in N_i(u)} r_{v,i}. \tag{2.1}$$

这个式子并没有考虑到邻居间可以具有不同相似度的问题，一个常见的解决方式是利用每一个邻居与 u 的相似度对评分加权。然而，如果这些权重的和不等于 1，那么预测的等级可能超出允许的范围，因此通常使用加权平均的方式进行计算。

2.1.2 基于物品的协同过滤

比较流行的是基于物品的方法 [6],基于用户的推荐依赖"志同道合"用户的观点,而基于物品的方法着重于相似物品的评分。为了使用相似性度量,首先确定 u 选择过的与 i 最相似的 k 个物品,这 k 个邻域的集合由 $N_u(i)$ 表示,预测值 r_{ui} 的计算方法是采用 u 在集合 $N_u(i)$ 中评分的加权平均:

$$\hat{r_{ui}} = \frac{\sum\limits_{j \in N_u(i)} w_{i,j} r_{u,j}}{\sum\limits_{j \in N_u(i)} w_{i,j}}. \tag{2.2}$$

2.1.3 相似性权重计算

构建基于邻域的推荐系统的最关键方面之一是相似性权重的计算,它对推荐系统的准确性和性能具有显著影响。相似性的计算基于这样的共识:相似的用户喜欢相似的物品,同时相似的物品被相似的用户喜欢。目前有多种方式用于相似性的计算,最常见的是将评分矩阵中的行列向量作为对应用户或物品的抽象,然后计算向量余弦夹角(以计算物品间相似性为例):

$$sim(i,j) = \frac{\sum\limits_{u \in U} r_{u,i} r_{u,j}}{\sqrt{\sum\limits_{u \in U} r_{u,i}^2} \sqrt{\sum\limits_{u \in U} {r_{u,j}}^2}}. \tag{2.3}$$

实际上,当使用显式评分作为偏好度时,不同的用户往往会有差异,例如某些用户的评分普遍偏高,可以使用与评分平均值的偏移作为偏好度,因此,采用这种调整的余弦相似度,物品 i 和物品 j 之间的相似度计算方法如下:

$$sim(i,j) = \frac{\sum\limits_{u \in U} (r_{u,i} - \bar{r_u})(r_{u,j} - \bar{r_u})}{\sqrt{\sum\limits_{u \in U} (r_{u,i} - \bar{r_u})^2} \sqrt{\sum\limits_{u \in U} (r_{u,j} - \bar{r_u})^2}}, \tag{2.4}$$

其中,r_u 是用户 u 评分的平均值。

2.1.4 基于用户与物品方法对比

应用基于邻域的协同过滤系统时,可以从以下几个方面考虑选择基于用户还是基于物品的方法 [5]:

1. 准确性:在电子商务系统中,用户的数量往往远大于商品的数量,数据的稀疏性会导致很难匹配到相似用户,因此推荐的精确性将收到严重影响,此时使用基于物品的方法会有更高的准确性,因为少量高置信度的邻居要比大量不那么

相似的邻居好得多。同样的，在用户数量少于物品的场景下，例如学术论文推荐系统，采用基于用户的方法会比较准确。

2. 效率: 推荐算法的计算量和存储量也取决于用户和物品的比例，当用户量远超物品数量，计算用户之间的相似性会产生庞大的计算量，影响系统的可扩展性。因为用户只对少数物品评级，因此仅存储非零的或者前 N 个相似性权重可以降低存储量和在线推荐的复杂度。

3. 稳定性:基于用户和基于物品的方法之间的选择还要取决于系统中用户和物品变化的频率和数量，一般情况下系统中可用的物品与用户相比是更加静态的，因此物品相似性权重可以不用频繁地计算，同时仍然能够向新用户推荐，这样，利用用户当前的数据可以进行实时的查询推荐，相比于基于用户的方法，系统具有更高的稳定性。

4. 可辨识性:面向物品的方法更适合解释推荐背后的原理，这是因为用户往往熟悉之前选择过的物品，而不知道那些所谓志同道合的用户。这样，可以向用户展示在当前预测中使用的邻居物品列表以及它们的相似性权重，通过修改列表或权重，用户可以交互的参与推荐过程。

2.2 基于潜在特征的协同过滤

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。

2.2.1 基本的矩阵分解模型

奇异值分解 [7] 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：

$$\min_{q^*,p^*} \sum_{(u,i)\in} \left(r_{u,i} - q_i^T p_u\right)^2 + \lambda(||q_i||^2 + ||p_u||^2), \tag{2.5}$$

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。

2.2.2 随机梯度下降

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：

$$\begin{aligned} q_i &\leftarrow q_i + \cdot(e_{ui} \cdot p_u - \lambda \cdot q_i) \\ p_u &\leftarrow p_u + \cdot(e_{ui} \cdot q_i - \lambda \cdot p_u) \end{aligned} \tag{2.6}$$

2.2.3 交替最小二乘

如果我们固定正则平方误差式子中的一个未知项，那么问题就转化成了二次函数求最值的问题。因此，交替最小二乘方法交替的固定 q_i 和 p_u ，当所有的 p_u 被固定，系统利用最小二乘法重新计算 q_i 的值，反之亦然。这个方法确保每一次都使之下降直至收敛。

通常情况下，随机梯度下降比交替最小二乘要简单也更快，但是在允许并行的系统下，交替最小二乘可以表现出很强的并行性，因为系统计算每一个 q_i 是独立于其它物品向量计算的，并且计算每一个 p_u 也是独立于其它用户向量的。另一个方面，当基于隐式反馈的数据时，训练矩阵中样本不再稀疏，如果使用随机梯度下降遍历所有样本项将是不可行的。

2.2.4 加入偏移量

基于矩阵分解方法的协同过滤的优点是可以很方便地处理不同种类的数据和一些应用中特定的需求 [8]，这需要在之前的学习框架上作出调整。观察到的样本数据

中通常会存在用户和物品个体性的偏移，例如一些用户给分普遍偏高。因此，直接将 $q_i^T p_u$ 视为最后的评分值是不明智的，系统尝试为每个用户和物品设置偏移量，并将偏移量的近似值加入到 r_{ui} ：

$$b_{ui} = \mu + b_i + b_u,$$

其中，μ 是全部评分的平均值，参数 b_i 和 b_u 是在样本上用户 u 和物品 i 相对于平均值的偏移量，加入偏移后改写为如下形式：

$$\hat{r_{ui}} = \mu + b_i + b_u + q_i^T p_u.$$

这样，观察到的评级被分为了四个组成部分：全局平均、用户偏移、物品偏移、用户物品匹配。这使得每一个部分可以单独解释其含义，系统通过最小化平方误差来学习模型参数：

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

2.2.5 概率矩阵分解

基本的矩阵分解算法如奇异值分解没有办法处理庞大的数据量，而且在矩阵很稀疏的情况下的表现也不理想。概率矩阵分解 [9] 为协同过滤的评分矩阵引入了一种概率模型的表示，它假设用户和物品的潜在特征向量均服从高斯分布，它们由如下过程取样产生：

1. 对于每一个用户 i ，选取用户特征向量 $u_i \sim N(0, {\lambda_u}^{-1} I_K)$.
2. 对于每一个物品 j ，选取物品特征向量 $v_j \sim N(0, {\lambda_v}^{-1} I_K)$.
3. 对于每一个用户物品对 (i, j) ，选取打分：

$$r_{ui} \sim N({u_i}^T v_j, {c_{i,j}}^{-1}).$$

其中，$c_{i,j}$ 作为正态分布的方差控制 r_{ui} 的准确度。

概率矩阵分解的复杂度与样本数呈线性关系，可以在庞大且稀疏的数据集上表现良好。

2.3 隐式反馈数据集上的协同过滤

推荐系统的任务是利用先前的用户反馈数据进行个性化建议来改善客户体验。系统依赖不同种类的输入数据，最方便的就是高质量的显式反馈，包括用户对产品的

兴趣的明确输入，例如，用户对物品的打分、或者仅仅是喜欢或不喜欢的标记。然而显式反馈并不可用，所以可以从更丰富的隐含反馈推断用户偏好，通过观察用户行为间接反映兴趣特征。

这些系统大量跟踪不同类型的用户行为，如购买记录、观看习惯和浏览时间，以模拟用户偏好。与显式反馈明显的不同，我们没有任何来自用户的关于他们的偏好的直接输入，特别是我们缺乏关于用户不喜欢什么物品的实质证据。我们把推断的用户物品偏好作为偏好度矩阵，其中的项 r_{ui} 可以是用户 u 购买物品 i 的次数，也可以是 u 浏览网页 i 的时间，找到合适的偏好度计算方式是基于隐式反馈推荐的关键。

Florham Park 等人提出了隐式反馈上的置信度模型 [10]，该方法将原始的观测值（r_{ui}）转换为两个独立的维度：偏好（p_{ui}）和置信等级（c_{ui}），这更好的反映了隐式反馈的特性。其中，用户 u 对物品 i 的偏好 p_{ui} 是通过二值化 r_{ui} 生成的：

$$p_{ui} = \begin{cases} 1 & r_{ui} > 0 \\ 0 & r_{ui} = 0 \end{cases} \tag{2.7}$$

也就是说，如果用户 u 访问过物品 i ，那我们可以说有迹象表明 u 喜欢 i ($p_{ui} = 1$)，另一方面，如果用户 u 从未访问过物品 i ，那我们觉得他不喜欢这个物品 ($p_{ui} = 0$)。但是我们信念应该与置信度相关联，因为零值并不一定意味着用户不喜欢这个物品，可能还有其他的一些原因，例如，用户可能不知道这个物品的存在或者由于其价格过高而不能消费它。另外，购买了一个物品也可能是很多不同因素的结果，例如，他可能购买后发现不喜欢这个物品。因此，我们推断用户偏好的项也有其置信等级。一般来说，随着 r_{ui} 增长，我们有更强的信心推断用户喜欢该物品，因此引入一组变量 c_{ui} 来衡量对推断的 p_{ui} 的信心，一种合理的选择是：

$$c_{ui} = 1 + \alpha r_{ui}$$

这样，对于每一个用户项目对，我们对 p_{ui} 有一些最小的信心，当我们观察到更多的正向偏好的证据时，我们对 $p_{ui} = 1$ 的信心就会增加，增加的速率由常数 α 控制。

我们的目标是为每个用户 u 找到一个向量 x_u ，以及为每一个物品 i 找到一个向量 y_i ，使得它们的内积 $p_{ui} = {x_u}^T y_i$ 表示用户对物品的偏好，本质上，这些向量尝试将用户和物品映射到共同的潜在特征空间，使得它们可以直接比较。这与基于显式反馈的矩阵分解技术相类似，但是有两个重要区别：(1) 我们需要考虑变化的置信等级。(2) 训练阶段要考虑所有的用户项目对，而不只是那些可见的样本。因此，通过

最小化如下的估价函数来计算模型的参数:

$$\min_{x^*,y^*} \sum_{(u,i)\in} c_{ui}(p_{u,i} - x_u^T y_i)^2 + \lambda(||x_u||^2 + ||y_i||^2), \tag{2.8}$$

$\lambda(||x_u||^2 + ||y_i||^2)$ 项用于正则化模型,以防止过拟合训练数据。λ 的值是根据不同的数据来确定的。

我们注意到估价函数中包含 $N * M$ 项,M 是用户数量,N 是物品数量,对于典型的数据集,$N * M$ 可以很轻松达到几十亿,这个庞大的数量限制了常见的显式反馈数据集的训练方法,例如随机梯度下降。利用变量的代数结构,可以通过线性时间内遍历所有用户项目对来最优化估价函数。

3 热门研究方向

3.1 混合推荐模型

我们知道数据稀疏性和冷启动问题是协同过滤方法的挑战,一种方法是建立完全基于用户和项目特征的预测模型,这种方法不会受到冷启动问题影响,在一些应用中,用户和项目都与一组信息特征相关联。例如,用户可以在注册时提供诸如年龄、性别、职业等个人信息,当项目是电影时,我们可以知道他们的类型、导演、演员等,对于新闻项目,我们可以从文章内容中提取特征。然而,这种完全基于内容的推荐不利用过去的交互数据,它也不能捕获存在于用户项目中的相关性 [11]。事实上,忽略特征而仅依赖用户与项目过去交互的协同过滤模型,对于旧用户和项目表现出良好的预测准确性。因此,最有吸引力的方式是混合过去的交互数据和特征,并且平滑地处理冷启动和热启动场景间的过度 [12]。

矩阵分解的一个优点是它允许在特征向量中加入附加信息。Wang and Blei [11] 提出了协同主题模型(CTM)用于学术文章的推荐,该方法使用两种类型的数据:用户的收藏历史和文章的内容,结合了基于潜在因素模型 [12, 9] 的协同过滤的思想和基于概率主题模型的内容分析 [13, 14]。对于每个用户,可以推荐类似用户收藏的旧文章和其内容反映用户的特定兴趣的新文章。潜在因素模型适合推荐已知文章,但不能推荐新加入的文章。为了推荐新文章,该算法使用主题模型,主题模型发现文章的潜在主题的表示,该组件可以推荐具有与用户喜欢的文章相似内容的其它文章,在没有评分的情况下,文章的主题表示使得算法可以对文章做出有意义的推荐。

3.1.1 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 [14]。最常见的主题模型是隐式狄利克雷分布（LDA）[13]，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 [13]。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。

3.1.2 协同主题回归

协同主题回归（CTR）模型结合了传统的协同过滤与主题模型，最简单的方法是直接使用主题分布表示可见的评分和单词，例如，我们可以使用主题分布 θ_j 替代公式（8）中的物品潜在向量 v_j ：

$$r_{ui} \sim N({p_u}^T \theta_j, {c_{i,j}}^{-1}).$$

这个模型的局限是它不能区分文档内容对不同用户的偏好，例如，两篇文档的主题分布相同，但是内容针对的用户群体不同。协同主题回归可以发现这种区别，该方法用对主题的兴趣表示用户，并且假设文档由主题模型生成。CTR 还包括一个隐式变量 ε_j 来调整主题分布 θ_j 在建模用户评分时的比例，预测时依赖内容和依赖协同过滤的比例由用户打分的数目来决定。CTR 的生成过程如下：

1. 对于每一个用户 u ，选取用户潜在向量 $u_i \sim N(0, {}_u{}^{-1} I_K)$.
2. 对于每一个物品 j ，
   (a) 选取主题分布 $\theta_j \sim Dirichlet()$.
   (b) 选取物品隐式偏移 $\varepsilon_j \sim N(0, {}_v{}^{-1} I_K)$ ，并且设置物品隐式向量为 $v_j = \theta_j + \varepsilon_j$.

(c) 对于文档中的每一个词 w_{jn} ：

i. 选取主题 $z_{jn} \sim Mult(\theta_j)$.

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

3. 对于每一个用户物品对 (i, j) ,选取打分：

$$r_{ui} \sim N(u_i{}^T v_j, c_{i,j}{}^{-1}).$$

CTR 的关键在于物品向量 v_j 如何生成,我们看到 $v_j = \theta_j + \varepsilon_j$,其中 $\varepsilon_j \sim N(0, {}_v{}^{-1} I_K)$,就等价于 $v_j \sim N(\theta_j, {}_v{}^{-1} I_K)$,因此物品向量 v_j 接近于主题分布 θ_j 。注意到 r_{ui} 的期望是 θ_j 的线性函数：

$$E[r_{ui}|u_i, \theta_j, \varepsilon_j] = u_i{}^T(\theta_j + \varepsilon_j).$$

因此这个模型被称为协同主题回归。CTR 模型很好的利用了内容信息,并将其结合到了传统的协同过滤算法中,使得系统在物品冷启动问题上表现良好。

3.2 跨域推荐系统

用户和物品的冷启动问题是推荐系统的固有限制,前文所述的 CTR 模型利用内容信息缓解了内容的冷启动问题,而对于新用户的冷启动问题往往不容易解决,因为新用户的信息通常不容易收集。为了解决它,跨域推荐系统利用辅助域中的用户反馈来协助目标域上的推荐任务 [15]。

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,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。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,例如,直观上来看,喜欢摇滚音乐的用户很可能喜欢科幻片,喜欢抒情音乐的用户或许喜欢爱情片。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 [16]。

Leizou [17] 提出跨域推荐的三个主要研究趋势:

1. 通过集成和利用分布在不同系统中的显式用户偏好。
2. 通过记录用户的行为和反应来描述用户特征,并利用这些信息生成多个域上的推荐。

3. 通过组合来自不同域的推荐来生成单一域上的推荐。

当两个域之间的用户和物品存在重叠时，我们可以直接将所有的信息看作属于一个共同的域，这样就可以利用传统的协同过滤进行跨域推荐。然而，当两个域没有重叠或重叠很小时，这种方法就会产生问题。为了解决上述不重叠的情况，我们必须找到某种方法，能够在域之间找到或建立某种类型的显式或隐式关系，其将被用作在推荐系统中连接不同域的语义桥 [16]。

跨域推荐的主要任务是在域之间找到或建立某种类型的关系的方法，这种联系将被用于连接不同域的语义桥。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，使得难以找到它们之间的关联。

3.2.1 迁移学习模型

一种方法是在两个不同的域中映射用户的潜在特征向量，对于一个用户 a ，假定他在目标域的潜在特征向量是 U_a ，辅助域中的为 U'_a ，我们的目标是找到它们之间的映射函数，使得它们可以相互转换以提高两个域的效果。直观上，理想的情况是找到 U_a 和 U'_a 之间的可逆映射函数，但是有时候这种关系是非线性的，在这种情况下不能找到可逆函数 [15]。因此，我们希望找到两个映射，$f(U'_a) \approx U_a$ 和 $g(U_a) \approx U'_a$ 。这样，一个域的用户特征向量可以被转换，用于推断另外一个域的用户特征向量。迁移学习是机器学习领域的热门研究课题，它的目标是通过利用其它域中已知的信息来提高一个特定域的学习效果。在迁移学习的框架下，学习过程的每一次迭代，先利用随机梯度下降等算法更新单个域中的参数，然后利用域间的映射函数进行估计并得到误差，以最大后验概率的方式调整映射函数的参数。实际中，为了降低噪声干扰和复杂性，通常只利用那些在两个域都有密集活动的用户集合来训练模型。

3.2.2 基于语义桥的跨域推荐

利用社交标签和语义信息可以建立不同域之间的桥梁 [18, 19]，因为不同域中使用的标签词汇之间的通常是重叠的 [1]，该方法的好处是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 [20]。

4 总结及展望

本文介绍了协同过滤推荐算法的主要思想及算法面临的主要问题，总结了近些年来的热门研究成果，讨论了针对数据稀疏性和冷启动问题的解决方案。

研究者提出了多种方法来解决协同过滤算法面临的数据稀疏性和冷启动问题，

一个明显的趋势是将其他领域的一些技术用到推荐系统中，例如利用迁移学习的方法实现跨域推荐模型，这些方法一定程度上缓解了这些问题，但是随着电子商务等平台的迅速发展，这些问题将进一步凸显，将领域外的一些技术与推荐算法结合起来解决面临的问题将是一个有意义的发展方向。

参 考 文 献

[1] 项亮. *推荐系统实践*. 人民邮电出版社, 2012.

[2] 王国霞 and 刘贺平. 个性化推荐系统综述. *计算机工程与应用*, 48(7):66–76, 2012.

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

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

[5] Christian Desrosiers and George Karypis. A Comprehensive Survey of Neighborhood-based Recommendation Methods. 2011.

[6] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In International Conference on World Wide Web, pages 285–295, 2001.

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

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

[9] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In International Conference on Machine Learning, pages 880–887, 2007.

[10] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit

feedback datasets. In Eighth IEEE International Conference on Data Mining, pages 263–272, 2008.

[11] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August, pages 448–456, 2011.

[12] Deepak Agarwal and Bee Chung Chen. Regression-based latent factor models. In ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 19–28, 2009.

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

[14] 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.

[15] Xin Xin, Zhirun Liu, Chin Yew Lin, Heyan Huang, Xiaochi Wei, and Ping Guo. Cross-domain collaborative filtering with review text. In International Conference on Artificial Intelligence, pages 1827–1833, 2015.

[16] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. Cross-domain recommender systems: A survey of the state of the art. 2012.

[17] Antonis Loizou. How to recommend music to film buffs: Enabling the provision of recommendations from multiple domains. University of Southampton, 2009.

[18] Manuel Enrich, Matthias Braunhofer, and Francesco Ricci. Cold-start management with cross-domain collaborative filtering and tags. 152:101–112, 2013.

[19] Chaochao Chen, Xiaolin Zheng, Yan Wang, Fuxing Hong, and Deren Chen. Capturing semantic correlation for item recommendation in tagging systems. In AAAI, pages 108–114, 2016.

[20] 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.