# 天津大学


**机器学习论文——浅析推荐算法在不同场景下的应用**

学　　院 智能与计算学部

专　　业 计算机科学与技术

年　　级 2021

姓　　名 黄琬婷

学　　号 3021244270

# 浅析推荐算法在不同场景下的应用

**摘要：**本文综述了各种推荐算法的原理和应用，为推荐系统在电子商务、社交媒体、音乐和视频流媒体等领域的实际应用提供了参考和指导。通过深入理解不同算法的特点，分析优缺点，可以更好地选择和优化推荐系统，满足用户个性化的需求。在具体应用方面，通过场景分析，指导了不同算法在特定场景下的选型，从而更好地匹配用户需求。以电影推荐系统为例，分析了用户评价数据、缺乏评价数据、关联性强的数据集等不同场景下的推荐算法选择和应用。同时，通过对推荐系统构建的步骤和实验结果的讨论，展示了一个简化但完整的推荐系统模型的构建过程。最后，对未来发展进行了展望，强调了基于深度学习的推荐算法的潜力。通过深入理解各种算法的原理，本文旨在为推荐系统的研究和应用提供全面的认识和指导。

**关键字：**推荐算法、协同过滤、SVD、基于内容算法

## 引言

推荐算法构成了推荐系统的核心，其基本思想包括通过对用户行为、项目特性等数据进行分析和深挖，以揭示用户的独特需求和兴趣，从而预测出用户可能会关注的项目。这些算法在电子商务、社交媒体、音乐和视频流媒体等领域得到广泛应用。

## 1 推荐算法的分类

### 1.1 基于行为的推荐算法

基于行为的算法，主要是经过协同过滤算法实现。协同过滤（Collabortive Filtering），通过用户和产品以及用户的偏好信息产生推荐产品的策略。基于领域的有两种：一类是以寻找具有相似兴趣的人群喜欢购买的商品为基础，也就是以用户为中心的推荐;另一类则是依据某人对某个特定商品的选择情况，进而向其推送与其相近或相关的其他商品，这就是以商品为中心的推荐方式。利用用户以及物品的信息来预测用户的喜好，并且发觉用户可能会喜欢的类似产品或者是喜欢产品的相关产品，这就是推荐系统的核心思想。而基于模型进行推荐的算法，

具有代表性的是 SVD 算法，它是一种隐语义模型的算法，其本质是对用户-商品数据产生的原始矩阵进行 SVD 分解，从而用更小的矩阵近似原始矩阵的过程。下面我们做具体介绍：

### 1.1.1 基于用户推荐算法

UserCF 的全称是 User-Based Collaborative Filter，基于用户的协同过滤算法。这个方法的基本思想首先是寻找"类似的家庭成员"，然后确定他们所喜爱的商品。简而言之，它依赖于用户之间的协作学习，即向用户推送其他与他们的爱好相近的用户喜欢的东西。

UserCF 算法主要包含了两个重要的步骤:

第 1 个步骤是，要找到与待推荐用户兴趣相似的用户集合。第 2 个步骤是，选出这些相似用户喜欢的，并且目标用户没有关注的物品，将它们推荐给目标用户。

实际上，其本质是求得不同用户集合的相似度，集合和集合之间的相似度计算方法，有很多种，比如有 Jaccard 相似系数、余弦相似度等等计算法方法。这里无非我们实现的是两者其一，我们以余弦相似度为例：

$$\mathrm{Cos(A,\ B)} = \frac{|A| \cap |B|}{\sqrt{|A| \times |B|}}$$

如果作为多维集合数据进行相似度计算，那么我们推演出的公式：

$$W_{AB} = \frac{|N(A)| \cap |N(B)|}{\sqrt{|N(A)| \times |N(B)|}}$$

得到用户之间的兴趣相似度后，我们要继续研究如何对某个用户进行推荐。

UserCF 算法会为用户推荐 K 个与他有相似兴趣的用户与商品。则获得公式为：

$$\mathrm{p(u,\ i)} = \sum_{v \in s(u,K) \cap N(i)} w_{rv} r_{vi}$$

p(u,i)是衡量某个用户对特定物品 i 的兴趣程度的指标。后面要计算出全部待推荐用户和待推荐物品的感兴趣程度 p。

### 1.1.2 SVD 算法

SVD（Singular Value Decomposition）奇异值分解，矩阵分解的一种方法。其算法本质是：对原始矩阵进行 SVD 分解，从而用更小的矩阵近似原始矩阵的过程。具体来说，假设，我们有一个可以通过特征值和特征向量求解矩阵特征的方

法，那么原始矩阵为：

$$A = Q \Sigma Q^{-1}$$

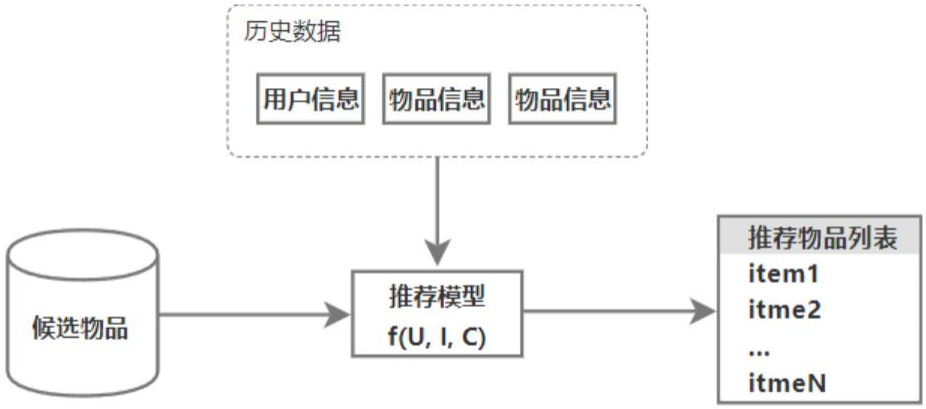
在这个过程中，Q 是一个 n×n 大小的矩阵，Σ则代表了由 n 个主导特征值构成的主对角线上的 n×n 矩阵。通常情况下，我们将 Q 中的 nn 个特征向量归一化处理，也就是让|wi|^2=1，这样一来，Q 的 nn 个特征向量就变成了标准的正交基，并且它们满足 Q^TQ=，换句话说就是 Q^T=Q^-1，这就意味着 Q 构成了一个酉矩阵。因此，我们可以得出以下结论：$A = Q \Sigma Q^{T}$。通过奇异值分解，我们可以满足对原矩阵分解，那么分解后的矩阵为：

$$A_{m*n} = U_{m*m} \Sigma_{m*n} V^{T}_{n*n}$$

在这个等式里，我们有 m×n 大小的方阵 A，它被称为左奇异向量和右奇异向量的集合体。Σ是一个 n×m 的矩阵，其特性在于非主对角线的所有元素都等于零，而主对角线上则被称作奇异值。同时，U 和 V 都被定义为酉矩阵，这意味着它们必须满足 U^TU=I 和 V^TV=I 条件。

### 1.1.3 ItmeCF 算法

基于物品的协同过滤算法是购物网站应用最多的算法。与基于用户的 CF 相比，它在计算过程中使用了物品间的相似性。也就是说，我们将所有用户对某个物品的喜好作为向量来计算相似性。得到相似的物品后，根据当前用户的偏好预测未来可能出现的物品。



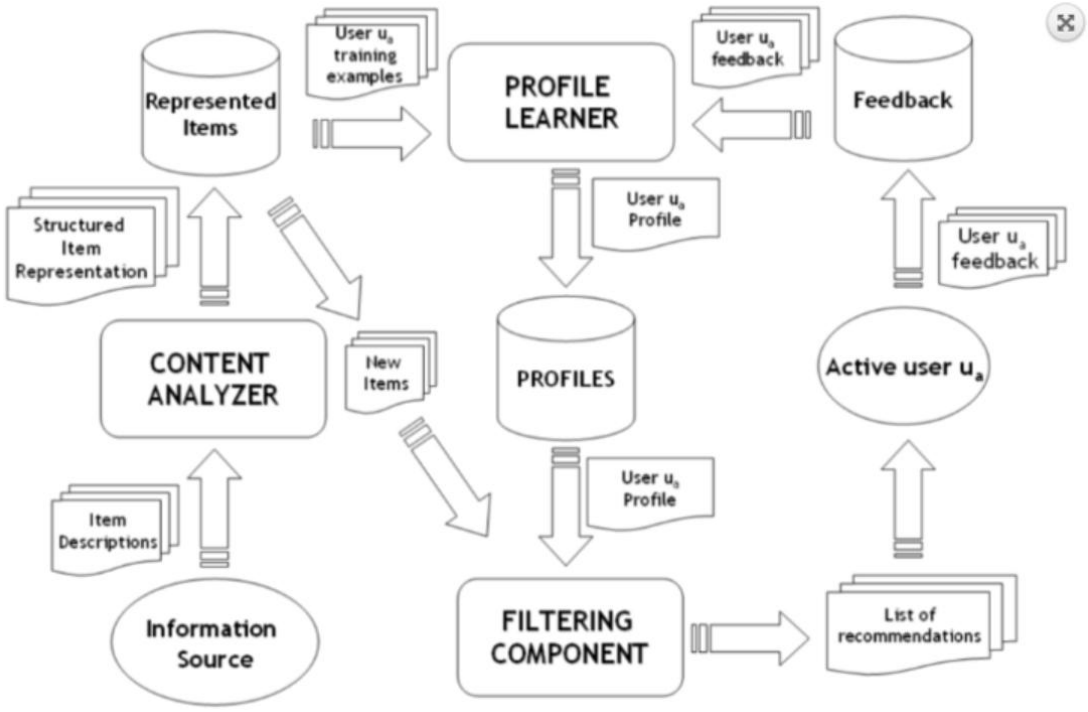
## 1.2 基于内容的推荐算法

协同过滤算法仅基于理解用户和物品之间的关系来进行推荐，而对物品本身的特性并不会有所考虑。然而，基于内容的算法则会对物品本身的特性进行考量。依据用户过去的购物经历，例如购买过什么商品、收藏了哪些商品、评价过哪些商品等，我们会根据这些数据来推荐与之相似的商品。

基于内容的推荐算法的原理如下：

选择一些具有象征意义的特性来描述每个事物。通过对用户历史行为数据的分析，我们可以理解物品的特性，进而掌握用户的喜好和兴趣，也就是构建出用户画像。

我们将对前一步获取的用户画像和待推荐产品的画像进行比较，这些画像是由待推荐产品的特性组成的。然后，我们会从相关性最强的前 k 个产品中选择目标用户未曾浏览过的产品，并向他们推荐。




## 1.3 基于规则的推荐算法

以关联规则为基石的推荐方法，其主要原理是通过已购买的产品来构建规则头部，并以此推测可能感兴趣的其他产品。这种方式能够揭示出各种产品的相互关系，并在零售领域取得了显著的效果。

关联规则的主要目标是在一个贸易信息系统中，计算出购买产品集 x 的贸易占比多大。这种方法直观地反映了消费者在选择某些产品时有多少倾向于选择其他类型的产品。

发现算法的初步相关规范是至关重要且耗费时间的，这也构成了计算的瓶颈。然而，它只能在线完成。另外，商品名称的同义性问题更是关联规则面临的一大挑战。

这种算法，如最多用户点击和最多用户浏览等，是一种常见的大众推荐方式，但在当前的大数据时代并不广泛应用。

我们通常会从用户购得的产品信息中寻找出现频率高的集合与顺序，以此进行高频组合分析，以确定符合支撑率门槛的相关产品的高频 n 个组件或系列。若用户已经购买过这些高频 n 个组件或系列中的某些商品，那幺我们将根据预设的标准对这些高频组件或系列的其他商品给予相应的评价并向他们提供建议，此标准可能包含支撑率、可信度及增益等因素。

### 1.4 基于知识的推荐算法

在某种程度上，基于知识的推荐可以被视为一种推理技巧，而不是根据用户需求和喜好来进行推荐。

首先需要构建用户知识与项目知识的基础框架，再根据明晰的需求来实施推荐。由于每个项目的系统都有各自的项目知识，因此它们的知识库也有所差异。利用知识驱动的推荐方式最大的优势在于无需过度依赖评级信息，并且不会出现冷启动的问题。

推荐的基础知识可以被划分为三种类型：基于 KDD（KnowledgeDiscoveryinDatabase）的推荐、基于 CBR 的 CaseBasedReasoning 推荐和基于知识推理的推荐。

KDD 基础上的推荐系统是将数据挖掘技术和传统推荐方式相融合，从中提取出隐藏在数据之下的有用知识与信息，进而为用户提供推荐。

CBR 推荐系统的基础是确定用户需求，然后筛选和选择已有的案例，并根据推荐的结果进行修正和调整。

推荐系统的构建依赖于知识推理，这包括用户、项目和功能三个方面。它们

通过一致的知识表达形式来展示用户概述和项目内容，并利用各种知识推理技巧在两者之间进行匹配，以此为基础向用户提供推荐服务。

### 1.5 基于上下文的推荐算法

众多的介绍系统主要关注用户和其他项目间的二元对立，却未考虑到日期、地址以及周围人等相关上下文讯息。将这些上下文信息加入推荐系统中能够显著提升其选择准确度。

将日期背景信息融入到推荐体系中，可以准确地展示使用者的变化兴趣，同时也能揭示项目的生命周期和季节性影响。当加入了时间内容后，推荐系统从静止转为移动，用户行为统计就会形成一段日期排序。

## 2 推荐算法的应用

### 2.1 算法的优缺点

| 推荐方法 | 优点 | 缺点 |
| --- | --- | --- |
| 基于内容推荐 | 推荐结果直观,容易解释；不需要领域知识 | 新用户问题；复杂属性不好处理；要有足够数据构造分类器 |
| 基于行为推荐 | 新异兴趣发现、不需要领域知识；随着时间推移性能提高；推荐个性化、自动化程度高；能处理复杂的非结构化对象 | 稀疏问题；可扩展性问题；新用户问题；质量取决于历史数据集；系统开始时推荐质量差； |
| 基于规则推荐 | 能发现新兴趣点；不要领域知识 | 规则抽取难、耗时；产品名同义性问题；个性化程度低； |
| 基于知识推荐 | 能把用户需求映射到产品上；能考虑非产品属性 | 知识难获得；推荐是静态的志扬工作室 |
| 基于上下文的推荐 | 提高推荐精度。 | 数据量大，计算复杂，算法运行效率低。稀疏性、冷启动、隐私与安全方面都存在问题。 |

表 1

## 2.2 不同场景下算法的应用

算法的运用往往与本身特点相关，其中，本身的优势可以作为某个特征点，匹配某个场景，那么我们可以基于场景进行算法的选型。

### 2.2.1 场景 1

数据包含用户评价等数据，并且拥有足够量的评价相关数据，那么大部分情况下，可以使用基于用户的推荐算法实现，通过评价数据建立其他的相似度的数据。例如：用户通常只是想看电影，但是并没有很明确的需求要看那部电影甚至是哪种类型的电影。从各大视频网站和评分网站的推荐理由来看，基于用户或者混合的推荐算法，即给用户推荐和他们曾经喜欢的电影相似的电影。

此外，亚马逊会根据用户之前的购物记录，比如曾经购买过武侠小说，继续推荐其他类型的武侠小说。

### 2.2.2 场景 2

几乎没有评价数据，并且用户结构单一，数据聚集类型明显，我们可以选择基于知识的推荐方式。例如，社区、兴趣圈的用户，往往通过基于知识（相同爱好）的算法推荐获得新的信息与新的好友推荐。

### 2.2.3 场景 3

拥有关联性很强的数据集求相似性的时候，可以以基于规则的推荐算法进行推荐。强调这种规则关联关系，往往能更准确的获得推荐结果。例如在银行客户交叉销售分析，这种场景中基于规则的算法被充分使用。

### 2.2.4 场景 4

电子商务中，基于物品的相似度，例如购买了该商品的用户还买了哪些商品。所以以基于物品的算法在物品数据集量很高的场景得以应用。

### 2.2.5 场景 5

大部分情况下，数据集往往数据量大，关联元素很复杂，这时候单一的推荐

算法无法有效或者轻易的获得推荐结果，我们就需要进行混合模式推荐算法。基于用户+基于内容的推荐算法或基于深度学习+基于内容的算法进行处理。例如电商，亚马逊、淘宝，都是运用多种算法混合实现对用户做产品推荐。

综上所述，不同算法有不同的应用之处，没有优劣之分。我们要巧妙利用，尤其是进行混合推荐之中，通过加权、切换、交叉、特征组合、串联层叠、特征补充、元级分层等技术，合理扬长避短来搭建更优化的推荐系统。

## 2.3 被动的选型及应用

算法本身也有某些局限性与优势，那么运用优势，规避局限性，从而解决或者避免某些问题，也成为了算法选型的标准之一。出现问题与缺陷时的选型应对：

**出现数据稀疏性问题**

在众多推荐系统中，所获得的评级信息相较于预期的项只占据了微小的比例，因此难以实现精确的模式匹配推荐。

利用用户个人信息以评估其相近程度。换句话说，如果两位用户不仅共享相似的影片评分，还可能是来自同一人口统计区域(例如，他们的性别、年龄、住址、受教育状况及职业等)，那么他们就被视为相似。此外，从顾客历史消费行为与反馈记录出发，寻找彼此之间的关联性也是一种方法。最后，我们运用降维技巧如奇异值分解 SVD，对稀疏矩阵进行降维处理。

**如何应对冷启动问题**

用户冷启动解决方案：使用基于人口统计学进行推荐；当评分数据足够多时基于用户协同过滤进行推荐。

物品冷启动解决方案：基于内容推荐；当收集到足量评分，基于物品的协同过滤推荐。

系统冷启动解决方案：引入专家知识，通过一定的高效方式迅速建立起物品的相关度表。

“探索与利用”机制：在“探索新数据”和“利用旧数据”之间进行平衡，使系统既能利用旧数据进行推荐，达到推荐系统的商业目标，又能高效地探索冷启动的物品是否是“优质”物品，使冷启动物品获得曝光的倾向，快速收集冷启

动数据。这里以最经典的探索与利用方法 UCB （Upper Confidence Bound,置信区间上界）讲解探索与利用的原理。

$$UCB(j) = \bar{x}_j + \sqrt{\frac{2\ln n}{n_j}}$$

其中 xj 为观测到的第 j 个物品的平均回报（这里的平均回报可以是点击率、转化率、播放率等），nj 为目前为止向用户曝光第 j 个物品的次数，n 为到目前为止曝光所有物品的次数之和。

通过简单计算可知，当物品的平均回报高时，UCB 的得分会高；同时，当物品的曝光次数低时，UCB 的得分也会高。也就是说，使用 UCB 方法进行推荐，推荐系统会倾向于推荐“效果好”或者“冷启动”的物品。那么，随着冷启动的物品有倾向性地被推荐，冷启动物品快速收集反馈数据，使之能够快速通过冷启动阶段。

## 3 推荐系统的构建

推荐系统的构建往往是很复杂，因为虽然场景的分析比较容易理清，但是因为用户数据量、数据集的稀疏，物品数据集的分布关系，机器性能瓶颈，系统内部实现架构复杂度等，往往很难准确的或者完整的实现某一种算法或者某几种算法。我们这里只取得最简化算法的模型进行解析与构建。我的实验即是完成的三种推荐算法的具体实现，而后构建出可应用系统模型。

我的实验主要实际应用三种不同的推荐算法，来实现电影推荐系统。以下三种为我实现的算法方案：

算法 1：基于用户相似度的协同过滤算法。该算法通过分析用户之间的相似性，将具有相似兴趣的用户推荐相似的信息或商品。这种方法的优势在于能够捕捉用户之间的行为模式，但也存在数据稀疏性和冷启动问题。

算法 2： 基于奇异值分解(SVD)的协同过滤算法。通过对用户-物品评分矩阵进行拆解，该方法有能力发现用户与物品之间的隐藏联系，从而提升推荐的精确度。然而，SVD 算法在处理大规模数据时可能面临计算复杂度较高的问题。

算法 3：协同过滤与基于内容的过滤相结（扩展以上代码以包含基于内容的特征）。以综合利用行为属性。这种方法可以弥补协同过滤在处理冷启动问题上的不足，提高推荐的覆盖性和精准度。

其内容主要包括：数据集分析、图表绘制、KNN 和 SVD 的性能对比、SVD 算法推荐等步骤，这一系列步骤构建了一个完整的推荐系统，为用户提供个性化的推荐服务。

这里我用基于用户的推荐算法讲解一下我的实现设想与步骤:

用户相似度的协同过滤算法通过分析用户行为历史和相似用户之间的关系来为用户推荐物品。其原理在于，如果两个用户在过去的喜好或行为上相似，那么他们在未来可能也会对相似的物品表现出兴趣。以下是主要步骤：

1. 数据收集：首先，需要搜集用户的行为数据，例如用户对物品的评分。通过这些数据构建一个用户-物品评分矩阵，其中行代表用户，列代表物品，每个单元格存储用户对物品的评分信息。

2. 用户相似度计算：为了找到相似的用户，需要计算用户之间的相似度。常用的方法包括余弦相似度、皮尔逊相关系数等。相似度度量有助于确定哪些用户在兴趣和行为上更为接近。

3. 相似用户的选择：选择与目标用户最相似的一些用户，以建立相似用户的群体。这可以通过对用户相似度进行排序并选择前几名相似用户来实现。

4. 生成推荐列表：针对目标用户，系统会找到他最相似的用户群体，并向该用户推荐这些相似用户喜欢的物品，但目标用户尚未互动过的物品。

5. 评估和反馈：生成的推荐列表需要定期更新，以反映用户兴趣和行为的变化。同时，通过使用评估指标如准确率、召回率、F1 分数等来衡量推荐系统的性能。

实验的结果评估指标使用了精确率（Precision）和召回率（Recall）这两个关键指标，以及综合考虑了二者的 F1 分数。以下是对这些评估指标的重新组织说明：

精确率（Precision）：精确率是指在所有被推荐的物品中，用户实际感兴趣的物品所占的比例。精确率衡量了推荐系统的准确性。

召回率（Recall）：这个比率是指在所有用户真正感兴趣的商品中，被成功推荐出去的商品所占的百分比。召回比率反映了推荐系统找到了多少用户感兴趣的商品。

这两个指标通常是相互影响的，即提高精确率可能会导致召回率下降，反之亦然。因此，综合考虑精确率和召回率是评估推荐系统性能的重要步骤。

而 F1 分数综合考虑了模型的准确性和召回能力，特别适用于在类别不平衡的情况下平衡精确率和召回率。计算结果 f1 分数的结果是准确度和召回率的综合平均数，其计算结果区间在 0 到 1 左右。该值越大，代表模型性能越优秀。

根据以上原理，我构造了一个整体可应用模型，即通过输入用户 id、感兴趣的类别和评分数量阈值，便可以直接得出推荐结论。

## 4 未来展望

基于深度学习推荐算法

融合了深层次的学习模型到经典的数据分析和推理工具如内容驱动的方法或者协作模式来实现个性化的信息推送服务[1] 是一种常见的做法;另外一种方式则是利用非指导性的训练过程把不同的产品分组在一起以获得更清晰的产品类别划分结果 [2];而第三种选择则是在有明确目标导向的前提下采用引导式的培训流程为每个具体的项目打上标签以便更好地理解其属性并将其按照特定的标准加以区分化[3],这三种策略都是借助多种类型的复杂数学建模框架例如前馈型的多级感应系统、线圈状的人工智能网路或是回环型的自适应控制模块等等这些先进的技术手段从原始的大量输入资料里提炼出来具有代表性和可操作意义的信息元素作为后续决策的基础依据

此种方式主要应用于对图像、文字和声音等信息的处理。通过深度学习的手段，我们能够从文本中获取其风格、类别及特点等属性，进而完成个性化推荐。而对于如音乐播放器这类以音响为中心的软件来说，首先需要把音频资料转换成数码形式，然后利用深度学习技术，使用数字化的方式表达音频特性(例如柔和、嘻哈或古典等)，这样就可以形成用户的聆听习惯。借助深度学习推荐系统的最大优点在于它能适应各种不同类别的输入数据，并且都能从中提炼出特征，构建模型，因此具有多样性的推荐能力。然而要获得更为优秀的推荐结果，就必须花

费更多时间去优化模型。

基于深度学习的推荐系统中的常用神经网络如下:

深度学习中的核心组件包含了入射端口、滤波单元(即过滤部分)、缩放模块及全部的链接环节与出击终端;而其中的"filtering part" 和 "scaling module", 也就是我们所说的“feature extractor”。这种模型通常被应用于对图形信息的解析上,因此常常会根据历史项目的照片资料去推测并提供相似主题或色彩构架的项目建议给客户参考使用。

循环神经网络主要用来进行序列数据分析，例如语音识别。若想确保语义翻译的准确性，就需依赖前文提到的环境条件。因此，在解决这种问题时，循环神经网络显示出一些优势。

**参考文献：**

[1] Khan M M , Ibrahim R , Ghani I . Cross Domain Recommender Systems: A Systematic Literature Review[J]. Acm Computing Surveys, 2017, 50(3):1-34.

[2] Tang J , Hu X , Liu H. Social recommendation: a review[J]. Social Network Analysis&Mining, 2013.

[3] Chen L , Chen G , Wang F . Recommender systems based on user reviews[J]. User Modeling and User-Adapted Interaction, 2015.