arXiv:2302.02579v1 [cs.IR] 6 Feb 2023

推荐系统：入门指南

Pablo Castells[a] 和 Dietmar Jannach[ba]马德里自治大学（UAM）[1],西班牙，[b]克拉根福大学，奥地利
pablo.castells@uam.es, dietmar.jannach@aau.at

2023年2月 本稿是 R. Baeza-Yates 和 O. Alonso 即将在ACM出版社出版的《信息检索高级主题》一书中的章节预印本

[1]本手稿基于在 UAM 开展的工作，与亚马逊无关。

摘要

个性化推荐已成为现代在线服务的常见功能，涵盖大多数主流电子商务网站、媒体平台和社交网络。如今，由于其具有很高的实际相关性，推荐系统领域的研究比以往任何时候都更加蓬勃发展。然而，随着我们如今观察到的推荐系统新应用场景不断涌现，新的挑战也层出不穷，无论是在算法要求方面，还是在此类系统的评估方面。

本文首先概述推荐问题的传统表述，随后回顾用于物品检索和排序的经典算法范式，并详细说明如何评估这类系统。之后，我们将讨论推荐系统研究中的若干最新进展，包括基于会话的推荐、推荐系统中的偏差，以及推荐系统在实践中的影响与价值等问题。

目录

1 基本概念 **5**

1.1 引言.5

1.2 推荐任务.7 1.3 推荐算法.10 1.3.1 作为机器学习问题的推荐 . . . . . 10 1.3.2 根据输入数据对方法进行分类 11 1.3.3 协同过滤.12 1.3.4 学习排序.15 1.3.5 神经推荐.16 1.3.6 基于内容和混合推荐系统.18 1.3.7 讨论.22

1.4 推荐器系统的评估.23

1.4.1 在线评估.24 1.4.2 离线评估.25 1.4.3 离线数据.26 1.4.4 推荐任务与指标.28 1.4.5 超越准确率.31

2 近期主题 39

2.1 序列推荐与基于会话的推荐.39 2.1.1 问题定义与术语.39 2.1.2 序列推荐与基于会话的推荐算法 . . . 40 2.1.3 序列式与会话式推荐系统的评估 . 42 2.1.4 讨论与展望.46 2.2 推荐中的流行度、偏差与重复出现.46 2.2.1 偏差消除.47 2.2.2 理解偏差.47 2.2.3 反馈回路.48

2.3 推荐系统的影响与价值.49

2.3.1 理解人在环路中对推荐的影响.51 2.3.2 推荐系统的消费者价值与商业价值 53

2.4 总结与未来方向.56 2.4.1 总结.56

2.4.2 延伸阅读与未来方向.57

1 基本概念

1.1 导言

自动化且通常具有个性化的推荐如今在在线世界中无处不在。如今，每当我们上网时，很有可能很快就会收到关于更多值得购买的物品、可下载的热门应用，或值得探索的新音乐或艺术家的推荐。这些个性化建议由推荐系统提供；推荐系统是软件组件，基于统计学和机器学习模型，确定应向个体用户呈现的最合适物品。由于推荐系统在实践中的广泛应用，以及其为消费者和企业创造价值的能力已得到证明，推荐系统可以被视为人工智能最引人注目的成功案例之一。如今，得益于此类系统的实际相关性以及机器学习领域持续快速的发展，该领域的研究正比以往任何时候都更加蓬勃。

从历史上看，推荐系统领域与信息检索（IR）领域有着诸多联系。例如，从较大的集合中检索并排序一组物品，是推荐和搜索任务的核心。两个领域往往还依赖类似的算法和机器学习模型，并采用相同或相似的评估方法来评估这些模型的表现。然而，推荐与信息检索也有许多不同之处。例如，在推荐场景中，物品的选择通常基于个体用户画像，而不是交互式查询。因此，从信息检索的视角来看，推荐可以被视为一种基于隐式查询（也可能取决于用户上下文）进行项目检索的问题设定，而非基于显式查询。此外，推荐通常是一种推送式通信，而信息检索系统通常是响应式应用。

信息检索的根源最早可追溯至20世纪50年代 [Sanderson and Croft 2012]。相比之下，现代推荐系统的基础主要是在20世纪90年代奠定的。1992年，*Tapestry*——一种个性化电子邮件过滤系统的原型——在施乐帕洛阿尔托研究中心开发出来 [Goldberg et al. 1992]。该系统的一项主要创新理念是在过滤过程中利用他人的意见（评分），而“协同过滤”（CF）这一术语也随着该系统而广为流传。此后不久，几个研究小组着手将 Tapestry 的基于规则的方法自动化；1994年，影响深远的 *GroupLens* 系统被提出 [Resnicket al. 1994]。其主要理念是

该系统自动确定与用户志趣相投的“邻居”，并向用户推荐这些邻居评分较高的新闻物品。在这项早期研究中，推荐问题被操作化为一个“矩阵填充”问题：推荐算法的输入是用户-物品评分矩阵，目标是预测矩阵中的缺失项。这种研究操作化方式以及协同过滤技术的使用，至今仍在研究与实践中占据主导地位。然而，在过去近三十年间，人们为推荐问题设计或应用了更多复杂的机器学习算法，包括各种矩阵分解技术——这些技术最早也于20世纪90年代得到应用 [Billsus and Pazzani 1998]——以及最新的复杂深度学习技术，例如 [Liang et al. 2018]。

使用协同偏好信号的另一种方法，是查看文档的内容，以个性化的方式对其进行排序和过滤。相应的基于内容的 过滤技术也在20世纪90年代得到研究，通常被称为信息过滤或个性化信息过滤 [Foltz and Dumais 1992]。这些系统通常主要基于已知的信息检索技术，例如文档的 TF-IDF编码[1] 。纯粹的基于内容的推荐系统，其核心思想是向用户推荐与他们过去喜欢的物品相似的物品，也就是说，不考虑其他人的意见。不过如今，我们经常会看到实际应用中使用混合 系统，将不同的推荐技术与数据类型结合起来。此类系统的早期示例之一是*Fab*[Balabanovi´c and Shoham 1997],，这是一个用于斯坦福大学网页推荐的基于内容的协同过滤系统。

这里讨论的大多数早期系统都是作为研究原型开发的，主要用于学术环境。不过，在20世纪90年代结束之前，已有若干关于在电子商务或音乐等领域成功部署推荐技术的报告，例如 [Schafer et al. 1999]。后来，亚马逊可能是最早大规模依赖推荐技术的组织之一 [Linden et al. 2003]。如今，如上所述，许多在线服务都将推荐作为用户体验的核心，例如 Netflix [Gomez-Uribeand Hunt 2015],Spotify [Semerci et al. 2019], 或 YouTube [Covington et al. 2016]。多年来，文献中也记录了许多成功案例，展示了推荐系统的潜在商业价值，综述见 [Jannach and Jugovac 2019]。

总体而言，该领域已达到一定程度的成熟和标准化，尤其是在研究问题的操作化和评估方法方面。然而，推荐问题远未得到解决，因为新的应用场景不断涌现，而这些场景迄今尚未得到研究社区的充分关注。本文反映了这一现状，主要分为两部分。在第一部分，即第 1.2 至第 1.4 节中，我们概述推荐系统的基本概念

[1]在词频–逆文档频率（*TF-IDF*）编码中，文档表示为实数向量，其中每个元素代表一个词项，其值表示该词项在文档中的重要性。

以及推荐系统通常如何接受评估。随后，从第 2.1 节开始，我们将讨论推荐系统研究中当前和未来的一些议题。

1.2 推荐任务

任何推荐系统的主要计算任务[2] 都是确定在特定情境下向用户展示哪些物品。因此，在任何具体的应用场景中，在实施并评估算法之前，都必须先回答一个问题：应该依据哪些标准来选择和排序物品。从最一般的层面来说，任何推荐系统的设计目的，都是为一个或多个相关利益相关者创造某种价值或效用 [Jannach 和 Adomavicius 2016]。在实践中，创造价值的方式有很多。例如，从消费者的角度来看，人们通常认为推荐系统能够缓解信息过载问题。而从提供方的角度来看，提供个性化推荐则有助于提升与企业相关的关键绩效指标（KPIs），例如销售额或客户留存率。我们将在后文第 2.3 节更深入地讨论推荐系统的影响和价值等相关问题。

学术研究通常旨在抽象掉各个领域的具体细节、特定应用或推荐服务提供商的商业模式。因此，主要的计算任务往往也会以相应抽象的方式表述，通常是计算给定用户对各个物品的绝对或相对相关性的估计值。从一般层面来看，这个问题可以形式化如下，参见 [Adomavicius 和 Tuzhilin 2005]。给定

- 一个用户集合U,
- 一个物品集合I 以及
- 一个效用函数f：$U \times I \to R$，它将用户和物品映射为取自全序集R （例如非负实数）的效用值,

推荐$i' \in I$中的物品，使其效用函数值最大。更正式地说：

$$\forall u \in U, i'_u = \arg\max_{i \in I} f(u,i) \tag{1.1}$$

通常，我们感兴趣的是推荐不止一个物品。可以通过返回效用值最高的N 个物品来实现。根据这一定义，推荐系统的算法研究（见第1.3节）旨在定义或学习效用函数f，其中f 可以基于不同类型的附加信息。

在前文提到的 1994年 协同过滤 GroupLens 系统以及此后数不胜数的相关研究中，例如，效用函数 f 旨在返回一个评分预测。f 所使用的附加信息是已知评分构成的用户-物品矩阵。

[2]推荐系统的其他计算任务可能包括生成解释，或在交互式系统中确定下一个对话动作。

表 1.1 展示了这样一个评分矩阵的示例，其中表格突出显示的具体任务是为用户 $u1$ 提供推荐。为此，需要预测用户 $u1$ 对尚未评分的物品$i4$ 和 $i5$ 的评分。

GroupLens 系统中 f 的具体实现采用了最近邻方法。相比之下，在纯粹的基于内容的推荐方法中，f 并非基于用户-物品评分矩阵，而是基于个体用户可观察到的过去行为，以及有关 I 中物品的额外信息。最后，在混合方法中，f 的实现可能会同时考虑评分矩阵、物品元数据，以及其他信息来源，包括人口统计信息、过去评分之外观察到的其他各种类型的用户行为，或用户在社交网络中的嵌入情况；另见第 1.3.6 节。

	i1	**i2**	**i3**	**i4**	**i5**
u1	3	4	3	**?**	**?**
u2		4	3		5
u3		1	3	1	
u4	4		3	2	3
u5	3		3	2	

表 1.1 用户-物品评分矩阵。

协同过滤是文献中最常讨论的问题设定，大量文献致力于从用户-物品矩阵中通常带有噪声的数据 学习f。相应地，多年来，各种机器学习算法都被用于学习f。从历史上看，用户-物品评分矩阵被认为包含用户显式指定的物品评分。如今，研究更加关注只有隐式用户反馈可用的情形。在这种情况下，评分矩阵中的条目并非例如从一星到五星不等的数值，而是二元值。因此，矩阵中的正向 值通常以 1 表示，意味着用户过去曾浏览或购买过某件物品。最常见的情况是，隐式反馈数据集不包含负反馈值，也就是说，取值要么是 1，要么缺失。显然，过去购买过某件物品并不一定意味着用户喜欢它；隐式反馈，即观察到的用户过去行为，也可能比显式物品评分包含更多噪声。

原则上，各种信息都可以纳入效用函数 f。文献中特别关注的一类信息是上下文，它指的是进行推荐时所处的特定情境。当前上下文可能会显著影响某项推荐的实用性。例如，某家餐厅可能适合在夏季推荐，却不适合在冬季推荐。因此，将一年中的时间作为上下文加以考虑，可能至关重要。总体而言，与其他类型的侧信息不同的是，对于特定用户 u，物品 i 的效用可能因当前情境而异。因此，有一种提议是相应地扩展f 的函数签名，以便

明确体现这一方面 [Adomavicius et al. 2022]。相应地，我们可以有

$$f : U \times I \times C \to R$$

作为扩展的效用函数，其中C 表示应进行推荐的上下文。

通常，在文献中广泛使用能够进行单个相关性预测的效用函数，无论是否考虑上下文。然而，如今，在绝对尺度上预测单个评分 （正如 Netflix 奖竞赛中所做的那样）[Bennett 等人2007], 已不再被视为实践中最相关的问题。相反，人们更加关注创建建议排序列表，从而形成了*top-N* 推荐任务。从技术上讲，可以使用最初为评分预测问题设计的相同算法，并根据预测评分对物品进行排序。或者，也可以使用排序学习算法。这类算法不单独考虑可推荐物品，而是直接致力于优化排序。

由于问题设定发生了变化——创建物品的 top-N 列表，而不是进行逐点预测——可以定义一种形式如下的替代效用函数f :

$$f : U \times L^* \to R$$

其中L^* 是I的幂集的所有长度不超过N 的排列所组成的集合，另见 [Quadrana 等人 2018]。给定此函数后，我们推荐使效用值R最大化的物品列表。相应地，在设计算法时，问题在于定义或学习一个能够预测长度受限的物品排序列表的效用的函数。

关于效用这一概念，需要注意的是，到目前为止，我们并未对效用应如何定义或衡量作出任何假设。过去，从 GroupLens 系统开始，人们一直将评分视为效用代理指标。目标是预测用户会给尚未见过的物品打多少分，然后推荐预测评分最高的物品，假定这些物品最有用。不过请注意，我们的定义并不局限于消费者价值，效用函数也可能会考虑提供方利润 ，因此会从多个利益相关者的角度考量价值，另见[Abdollahpouri 等，2020a、Mehrotra 等，2018b]。

在传统场景中，主要考虑形式为 $f : U \times I \to R$ 的效用函数，其中每件物品的效用都独立于可能在同一推荐列表中展示给用户的其他物品。然而，这种方法无法评估整个前N 项物品列表的质量，例如无法衡量其多样性。因此，如今常用的效用函数及相应的评估指标会考虑点式效用估计以外的因素，另请参阅第 1.4 节关于推荐系统评估的内容。

1.3 推荐算法

推荐算法的发展自然地反映了任务定义的演变，同时也伴随着针对该任务设计的评估程序和指标（我们将在第 1.4 节中进一步讨论）。从本质上说，推荐可以作为一个监督学习问题来处理：给定已观察到的用户选择样例，我们旨在预测当前或未来（尚未观察到）的用户兴趣。任务表述的变化催生了不同的算法方法，而针对不同任务进行评估时，适用的指标也各不相同。

作为一个机器学习问题，推荐具有相当独特的性质。在这一领域，推荐之所以与众不同，本质上是因为推荐任务的核心涉及人的因素。在这些任务中，输入信号和预测目标的核心都由用户行为构成或与之相关。与例如识别图像中的形状，或根据医学检查诊断疾病相比，这带来了特定程度的复杂性。此外，推荐往往不只是预测人们的动作，还要通过让人们意识到可能更好的选择来改善（并因此改变）这些动作。因此，何为好的推荐（以及由此产生的何为好的算法）在很大程度上具有主观性，这是一个颇具挑战性的问题，我们将在第 1.4 节中进一步讨论。

传统上，推荐算法分为协同过滤 和基于内容的 [Adomavicius et al. 2022]。后者遵循这样的原则：人们的品味与物品固有的特征相关，并且往往会随着时间推移而保持稳定。前者则以多种不同方式，基于这样一种假设：用户偏好和选择在用户-物品空间中的分布存在规律和结构。例如，有些用户在个人兴趣方面可能与其他用户存在非随机的相似性。同样，这些结构也被假定会随时间推移而保持稳定。

不同的算法方法各有优势与不足。因此，稳健有效的推荐系统会结合不同方法中的元素，甚至采用其完整组件。为了使算法分类涵盖所有情况，这类复合解决方案通常被称为混合系统。与许多机器学习或信息检索系统一样，生产环境中的推荐系统通常会结合一种核心初始算法（通常是混合算法）来生成基础物品排序，随后再通过后处理算法、业务规则（例如实施营销约束）和精细优化，进一步提升最终推荐排名质量，并优化其他目标 [Amatriain 和 Basilico，2015，Covington et al. 2016]。

1.3.1 将推荐视为机器学习问题

现代推荐算法往往被明确表述为监督式机器学习任务：学习一个效用函数 f （采用前文上一节提到的某种形式），

使其最小化一个代价函数，而该代价函数的定义中包含效用函数。例如，代价可以是所有可用训练评分上的平方预测误差之和（按照监督机器学习术语，后者充当唱片公司），也可以是 Sigmoid 函数的二元交叉熵损失，或者是效用函数将物品对错误排序的数量（通过某种单调函数进行度量）。在这种表述下，各种推荐算法的区别体现在效用函数所采用的形式（参数化模型）、代价函数，以及最小化后者的过程上。

一个效用函数的例子是用户和物品嵌入 向量的点积[3]„，其中这些向量的坐标是待学习模型的参数，正如我们将在下文第 1.3.3.2 节中看到的那样。因此，该函数涉及某种任意选择（一个函数族，通常称为假设 空间）；学习该函数通常意味着通过将其参数拟合到训练数据（最小化训练标签上的代价函数）来学习这些参数（即“模型”）。不过，有些算法并不显式表示代价函数，而是通过启发式方法定义效用函数，而非学习得到。稍后在第 1.3.3.1 节中概述的常见最近邻协同过滤方法就是一个例子。无论哪种情况，效用函数都有需要优化的超参数，通常的做法是从训练数据中单独留出一个验证子集。所有算法的最终目标都是最大化某些关注的最终指标，例如精确率、召回率或收入——这就是为什么有些算法会将最终指标本身或其更易处理的近似值作为待优化的目标函数。

接下来，我们将概述推荐领域的主要算法方法和原理。首先简要讨论推荐所用的不同类型的输入数据，这些数据决定了基于其开发的算法方法之间的总体差异。每个算法家族都值得单独写一篇论文（甚至一本书）。因此，我们在此只提供概括性概述，并重点介绍一些主要内容。

1.3.2 根据输入数据对方法进行分类

推荐算法可以基于不同性质的输入数据。推荐系统最常见的数据来源——我们还可以说，这也是最能支持强大方法的数据来源——或许是记录下来的用户-物品交互：即前面讨论过的评分矩阵。如果算法的唯一输入就是这些数据，我们就说该算法采用协同过滤方法 [Adomavicius 和 Tuzhilin 2005]。协同过滤基于这样一个抽象原则：人们在做出未来选择时，不应只依靠自己的经验和发现，也可以从他人的经验和发现中受益。这一原则简单明了，而集体数据作为预测来源又具有巨大潜力；但这也正是纯粹采用协同过滤方法的致命弱点：当交互矩阵

[3]嵌入是对象的低维（稠密）学习表示，形式为实数向量，另请参见第 1.3.3.2 节。

稀疏时，由于缺少足够的输入信息，算法很难做出可靠的预测。在冷启动情形下，由于记录到的交互很少，协同技术可能无法提供恰当的推荐，例如新部署的系统刚开始运行时，或有新用户或新物品加入系统时。然而，当有足够的交互记录可用时，协同过滤是最有效、最强大的方法之一。

推荐算法除了用户-物品交互之外可以接收的所有其他输入，通常称为侧信息，参见第 1.3.6 节。使用这类信息是对协同过滤的良好补充，尤其适用于用户-物品矩阵中的稀疏区域。最常见的一类侧信息是任何可用的、与物品直接相关的数据和元数据，我们可以称之为物品特征：分类体系（例如电影类型、音乐风格、网店类别）、标签、与物品相关的自由文本（例如产品描述和评论）等。仅使用这类数据的算法称为基于内容的 [Pazzani 和 Billsus 2007]。纯粹的基于内容的方法不会向其他用户寻求有关客户可能喜欢什么的提示，而是一次只关注一个用户，考察目标用户的交互记录[4]——尤其是用户曾交互过的物品的特征，以及待推荐物品的特征。推荐系统还可以有效利用其他若干信息来源来提升推荐的效果和价值，例如用户之间的社交交互 [Guy 2022], 、用户相关数据（例如人口统计信息）以及上下文信息（例如用户和/或物品的地理位置、会话、时间、设备等）。

当有大量交互数据可用时，协同过滤作为基础算法核心可能比其他方法更有效；它不仅能提供准确的预测，还能为用户带来超越其个人先前经验的丰富选择 [Vargas 和 Castells 2011]。此外，这些算法具有相当强的通用性，不依赖特定领域，因为它们不对物品的类型作任何假设。

1.3.3 协同过滤

1.3.3.1 最近邻

最早的协同过滤算法受到一种典型的人类口碑行为的启发：人们在做决定时会向信任的朋友征求意见——这也是人们做决定和选择时获取指引的主要途径之一。遵循这一比喻，所谓的*k*近邻（kNN）算法会根据目标用户朋友的建议加权平均来推荐选择。

根据这一启发式直觉，推荐可以视为一个回归问题，其中 k近邻算法将用户 u 对物品 i 的偏好程度预测为一组经过筛选的用户对*i*的观测偏好的线性组合；这组用户是目标用户的

[4]“目标用户”一词通常用来指我们向其提供推荐的用户。

邻居$N[u]$。沿用上一节的符号，物品根据以下效用函数进行排序：

$$f(u,i) = C_i \sum_{\substack{v \in N[u] \\ r(v,i) \neq \emptyset}} w_{u,v}\, r(v,i) \tag{1.2}$$

其中C_i是一个归一化项，而$r(v, i) \neq \emptyset$表示涉及v和i的交互数据（“评分”）可用。线性组合中权重的典型定义基于成对用户相似度$w_{u,v} = \mathrm{sim}(u, s)$，其中相似度函数可以是余弦相似度，或用户$u$和$v$的评分向量之间的皮尔逊相关系数。为了处理用户向量中的缺失值，这些度量的标准定义需要进一步具体化。例如，使用余弦相似度时，缺失评分可以赋值为零。

邻域$N[u]$通常选取与u最相似的前k个用户，其中k成为该算法的一个超参数。尽管如此，邻域选择是一项模块化操作，已有多种变体和改进。当推荐被视为评分预测时，上文公式 1.2 中的 k近邻得分会除以$C_i = 1/\sum_{v \in N[u]: r(v,i) \neq \emptyset} |\mathrm{sim}(u, v)|$进行归一化，以便生成评分尺度上的数值。当评估任务是排序任务时，研究发现省略归一化（即$C_i = 1$）往往能取得更好的结果 [Cremonesi 等人 2010]。

由此衍生出许多其他 k近邻方案变体，例如广泛使用的基于物品的版本（大致上是交换用户和物品的角色）、多种相似度函数、不同的邻居选择方法等。尽管 k近邻最初是作为一种启发式方案开发的，Ca˜namares 和 Castells [Ca˜namares 和 Castells 2017]证明，k近邻可以通过概率形式化来描述，其中排序函数由一个需要最大化的目标函数推导得出：用户会喜欢的推荐物品的预期数量。关于 k近邻变体的更全面综述，请参阅 [Nikolakopoulos 等人2022]。

k近邻是一个由来已久的想法 [Linden et al. 2003]，但如今仍是一种颇具竞争力的方法 [Ferrari Dacrema et al. 2021，Ludewig 等人， 2021]。与其他较新且更复杂的方法相比，它易于理解、实现简单，计算成本也较低。因此，在实验性推荐系统研究中，将其作为参考基线纳入研究是值得推荐的做法 [Ca˜namares 和 Castells 2018，Cremonesi 等人， 2010]。

1.3.3.2 矩阵分解

到 2000 年代中期，矩阵分解在该领域开始流行，并很快成为首选的协同过滤方法，这得益于它在实证中的效果、相对简单的特点，以及作为框架的灵活性，能够支持多种算法开发与变体。矩阵分解基于这样一种假设：用户的兴趣可以用一个低维的潜在因子空间来描述，该空间综合了物品的底层属性

，这些属性决定了人们可能喜欢这些物品的原因 [Koren 等人， 2009]。按照这一观点，用户和物品都可以表示为共同潜在因子向量空间中的向量；具有相似品味的用户会拥有相似的因子向量，同样，能取悦相似用户的物品也会如此。近年文献中，这些低维向量被称为嵌入，由此与自然语言处理和信息检索等领域中的类似技术建立了联系 [Mikolov 等人，2013，Pennington 等人。 2014]。

潜在因子不一定对应于物品的实际属性或用户特征，也未必是我们按照人类认知习惯所能描述的属性。算法方法将它们视为抽象维度，并为用户和物品赋予相应的数值；这些维度通常无法通过目测或明确的自然直觉来解释。矩阵分解算法只需确定我们希望考虑多少个因子——这会成为该方法的一个超参数——而不必确定这些因子具体是什么，只需将它们视为多维向量空间中的坐标轴即可。

形式上，对于一个k维因子空间，用户$u \in U$和物品$i \in I$分别用向量$p_u \in \mathbb{R}^k$、$q_i \in \mathbb{R}^k$表示，其中坐标$p_{u,z} \in \mathbb{R}$、$p_{i,z} \in \mathbb{R}$分别表示用户u对因子z的兴趣程度，以及因子z在物品i中的含量——z坐标值越高，用户对该因子的兴趣就越浓，该因子在物品本质中的重要性也越高。基于这种表示方式，我们还作出一项假设：用户u对物品i的兴趣可以通过它们的潜在向量的点积来表示。按照效用函数的记号，$f(u,i) = q_i^t p_u$。

生成潜在向量的常见方法是最小化代价函数$\mathcal{R}$（在经验风险最小化机器学习框架中也称为风险或期望损失 [Vapnik 1998]），该函数通常包含对训练数据进行预测时产生的误差，并加入一定的正则化。例如，一种基本的公式为：

$$\mathcal{R}(p,q;\lambda) = \sum_{u,i}(r(u,i) - q_i^t p_u)^2 + \lambda(\|p\| + \|q\|) \tag{1.3}$$

平方差上增加的这一项是一种常见的正则化因子，用于避免过拟合，其中 $\|\cdot\|$ 表示 L_2 范数， λ 是通过交叉验证调整的超参数。通常使用随机梯度下降或交替最小二乘来求解这一最小化问题，从而得到易于实现的方法 [Koren 等人 2009]。算法学习到潜在向量后，点积 $f(u,i) = q_i^t p_u$ 便用作推荐的排序函数。

代价函数中的求和可以仅针对有训练数据的用户-物品对进行，但通常会将求和范围扩展到所有未观测到的用户-物品偏好，或其中的一个样本子集 [Koren 等人 2009]。在后一种情况下，需要为缺失的偏好观测值插补一个值，插补方式可以有多种（例如使用常数参数值或随机值）。方程中的误差项

1.3 可以针对每个用户-物品对采用不同的权重，通常用于表示对应偏好数据点（观测到的或插补得到的）的置信度。最常见的做法是，对有训练数据和没有训练数据的用户-物品对分别使用两种不同的权重值 [Hu 等人。 2008]误差加权与缺失值插补相结合，体现了推荐任务中的不同细节，并会导致不同的算法行为（例如参见 [Steck 2013]）。除这些配置选项外，矩阵分解还衍生出许多其他变体，例如将物品、用户和全局偏置项作为额外参数加入评分表示中 $q_i^t p_u$ [Koren 等人 2009], 时间感知 [Koren 2009], 概率形式 [Hofmann 2004年、Salakhutdinov 和 Mnih 2007], , 以及许多其他扩展。尤其是，代价函数的各种变化催生了全新的方法，接下来我们将对其进行概述。

1.3.4 学习排序

随着人们认识到，在现实场景中，推荐效果更多取决于物品排序，而非逐点评分预测，该领域开始转向将推荐视为学习排序（LTR）问题，这一趋势借鉴了信息检索领域更早出现的类似发展 [Joachims 2002]。广义而言，学习排序会引入损失函数，明确考量物品排序，而非评分误差或交互预测误差 [Rendle 等人 2009]。在各种不同的实现方式中，代价函数通常表示待学习模型生成的排序结果与依据训练偏好值应有的物品排序之间的“矛盾”程度。换言之，模型经过训练后，会尽可能提高其排序与现有观测的一致性。大多数方法采用成对方法，将“排序误差”定义为物品对排序错误的程度：

$$\mathcal{R}(\theta) = \sum_u \sum_i \sum_j \ell(u, i, j|\theta)$$

其中， θ 是推荐模型的参数，而损失 $\ell(u,i,j|\theta)$用于量化模型 θ 对 u 的i 和j 进行排序时，与现有观测相矛盾的程度（并且ℓ 涉及待学习的效用函数 f——即模型）。对 u 拥有训练数据的所有物品对i, j 求和即可。

该领域一个具有代表性且有效的例子是贝叶斯个性化排序（BPR） [Rendle 等人 2009], 提出的方法，它已成为文献中常见的参考方法和基线。BPR 中的排序不一致性通过如下概率来表述和推导：学习到的模型所预测的分数 θ 对物品对进行排序时，与训练评分数据相矛盾的概率。概括而言，省略细节后， $\ell(u, i, j) = \mathbb{1}_{r(u,i)>r(u,j)} P(j \text{对于} u \text{的排序高于} i \,|\, \theta) P(\theta)$。排序先后关系的概率通过将排序分数差表示为逻辑西格莫伊德函数来平滑处理（以确保可微性）。

例如，对于一个矩阵分解模型，其中 θ 由潜在向量 p_u、q_i 构成，适用于 $(u, i) \in U \times I$，且效用评分函数为 $f(u,i) = q_i^t p_u$，我们得到 P(j 对于 u 的排序高于 i $|$ p,q) : $= 1/(1 + e^{q_j^t} pu - qti\ pu)$。在正态性和因子独立性假设下，先验 P(p,q) 会产生典型的 L_2 正则化项。随后，可通过随机梯度下降最小化所得代价函数。因此，BPR 可以视为矩阵分解之上的一层，也是训练潜在向量的一种方法，与逐点（即逐用户-物品）代价最小化不同。

基于类似原理，人们提出了许多替代这类方案的方法。例如，同样采用成对方法的 RankALS [Takˊacs and Tikk 2012] 本质上将 $\ell(u, i, j) = ((r(u, i) - r(u, j)) - (q_i^t p_u - q_j^t p_u))^2$ 作为需要最小化的核心成对排序误差，并采用交替最小二乘而非梯度下降作为最小化过程。其他方法则将目标排序评估指标（例如归一化折损累计增益、平均准确率均值（MAP）或平均倒数排名（MRR））引入目标函数；也就是说，目标函数近似衡量物品排序不理想时，该指标会损失多少。例如，CLiMF [Shi 等人 2012] 将目标函数定义为 MRR 的平滑下界近似。在学习排序文献中，这些方法被称为列表式[Liu 2009], ，因为尽管损失函数通常仍会转化为成对形式，但它反映的是排序优劣函数（目标评估指标），而不是与指标无关的二元排序中的物品对分类误差。

1.3.5 神经推荐

矩阵分解将潜在向量的点积视为评分函数，而在推荐中，也可以考虑更复杂、未必是线性的[5] 效用评分。一种可能的方法是使用神经网络这一类非线性函数；这类函数的性质和训练技术已得到深入研究 [He 等人，2017，Zhang 等人， 2022]。深度学习在计算机视觉、语音识别和自然语言处理等机器学习应用领域掀起变革后，近年来作为设计各种推荐方法的基础而迅速走红。全面介绍这一领域超出了本文的范围，因此我们仅作概述；如需更全面的综述，读者可参阅例如 [Zhang 等人， 2022] 。

有几个技术和战略方面的原因，让我们对将深度学习用于推荐这一方法持乐观态度：能够逼近任意预测函数；训练效率高（与复杂度相当的其他机器学习方法相比）；能够顺畅整合异构数据源（包括非结构化内容）、数据维度和预测信号；减少特征工程的工作量；已在其他机器学习领域证明了实证效果；以及软件资源日益丰富

[5]严格来说，早期的矩阵分解模型应称为双线性模型：评分函数是关于模型参数的二次多项式（因此并非真正的线性函数），但分别关于用户参数（潜在因子）和物品参数是线性的——这就是双线性的含义。

以及从深度学习在众多领域蓬勃发展的实践中积累的知识。此外，作为通用函数逼近器，神经网络可以视为几乎任何更简单模型（例如前文所述矩阵分解中的点积）的平滑泛化。从这个角度来看，严格来说，神经网络与非神经网络之间的二分法或许已无关紧要。然而，深度学习中的“深度”表明，人们正通过“更大”的分层模型，切实利用扩展模型和处理复杂性的机会。

神经网络能够提供任意高的表达能力，以捕捉用户-物品空间中的复杂关系，而矩阵分解模型可能无法掌握这些关系。在顶层（输出层），损失函数可以采用逐点（评分或二元预测误差） [He 等人2017] 或成对（排序误差）元素 [Song 等人 2018] ，这与前文讨论的其他矩阵分解和学习排序方法类似。神经网络架构的通用性还使其能够在用户-物品交互之外，顺畅地整合侧信息。我们将在第 1.3.6.3 节更广泛地讨论侧信息在推荐中的使用。该领域能够发展出的各种扩展和组合，几乎可以与想象力一样广阔。文献中已经提出了大量复杂的网络架构，采用了自编码器 [Liang et al. 2018], 、卷积神经网络 [Tang 和 Wang，2018，Yuan 等人 2019], 、循环神经网络 [Hi-dasi 和 Karatzoglou，2018，Wu et al. 2017], 以及注意力层 [Kang 和 McAuley 2018],等网络结构，这里仅举几例。针对特定任务，人们还开发了专门的架构，例如将在第 2.1 节讨论的基于会话的推荐。有趣的是，推荐领域的最佳性能有时是由比其他领域更浅层的神经模型取得的 [Anelli 等人，2022，Steck 2019]。

深度学习已被证明在推荐领域具有实证效果，产业界也已开始采用这项技术 [Covington 等人，2016，Okura 等人，2017，Wu 和 Grbovic 2020]。这一领域的某种炒作，也可能导致早期实证分析中出现了一定程度的噪声和仓促之处，有时让人难以准确判断神经网络方法的实际效果，或确定它们在哪些条件下最适用 [Ferrari Dacrema et al. 2021，Garg 等人，2019，Kouki 等人，2020，Ludewig 等人，2021，Rendle 等人，2020，Steck 2019]。在序列推荐等特定任务中 [Sun 等人，2019b], ，以及在具备侧信息或海量数据等特定条件下，已发现明确的效果提升。深度学习通常还能简化特征工程，减少相关工作量。有关 Netflix 推荐系统中深度学习的讨论，请参阅 [Steck 等人 2021] 。

深度学习在推荐领域面临着与其他领域类似的挑战：需要调优大量超参数，需要在达到效果潜力的同时权衡高昂的训练成本，需要拥有海量数据，而且其黑箱特性还带来了可解释性方面的挑战 [Afchar 和 Hennequin 2020]。因此，神经网络方法的成本和复杂性未必总能转化为效果提升，因而不一定适合作为解决所有推荐场景和问题的通用方案 [Ferrari

Dacrema 等人，2021年、Rendle 等人 2020]。另一方面，充分发挥深度推荐架构的作用通常需要丰富且大量的数据，而这对于企业边界之外的研究而言并非总是可得。尽管如此，神经推荐仍是一个蓬勃发展的领域，目前顶级研究出版物中正持续发表大量相关工作，产业界也在努力挖掘其潜力 [Cen 等人，2020年、Covington 等人，2016、Feng 等人，2020年、Steck 等人 2021]。我们建议读者继续关注近期文献和综述 [Zhang 等人 2022] ，并持续留意进行中的研究、未来可能出现的突破。

1.3.6 基于内容的推荐系统与混合推荐系统

正如上一节所述，协同过滤（CF）方法能够检测并利用用户群体中的偏好或行为模式，因此在许多应用场景中都非常有效。此外，它们尤其擅长帮助用户发现此前不熟悉的内容类型，例如参考志趣相投的用户的偏好。协同过滤方法的一个引人之处在于，即使完全不了解物品本身（例如某件物品在电子商务商店中的类别），它们也能实现这一点。

不过，如果我们掌握了有关物品的知识，以及用户对这些物品的特定属性的偏好，那么在推荐过程中考虑这些信息自然也很有意义。例如，如果我们知道某位用户通常偏好某些电影类型或新闻类型，那么推荐更多同类内容是顺理成章的。此外，如果我们尚无已记录的交易或评分数据，也就是说，用户还没有为许多物品评分（用户冷启动），或者目录中有尚无购买历史的新物品（物品冷启动），协同过滤方法可能无法很好地发挥作用。

1.3.6.1 纯基于内容的系统

在这类情况下，当可用的协同信息很少或完全没有时，一种做法是将用户个人 过去的偏好与有关物品特征的知识相匹配。传统上，这一过程称为基于内容的过滤[6]。从技术上讲，许多早期的基于内容的过滤方法都基于信息检索领域的理念和技术。在推荐和检索场景中，目标都是识别并排列一组相关文档。在信息检索的情形中，起点是一个（搜索）查询；而在推荐系统中，检索过程则基于基于内容的用户画像。

[6]早期推荐系统的一个重要应用场景，是对收到的新闻消息进行个性化筛选；这些应用中的筛选依据是文档内容，也就是文本文件。后来，基于内容的 一词也用于指依赖物品元数据（例如类型）的系统。关于这一主题的早期综述，请参见 [Pazzani 和 Billsus 2007] 。

从一般层面来说，基于内容的推荐可以概括如下 [Adomavicius 和Tuzhilin 2005]:

$$f(u,i) = score(ContentBasedProfile(u,Content(i)) \tag{1.4}$$

给定用户u 和物品i 的预测相关性f，是基于评分函数$score$ 将u 的基于内容的用户画像与物品i的内容特征进行匹配得出的。

实现这种基于内容的方法有多种方式。在物品i 有文本描述可用的情况下，传统做法是使用 TF-IDF（词频-逆文档频率）编码来表示物品。在这种实现中，函数 *Content(i)* 会返回一个实数向量，其中每个元素对应一个词项（单词），其数值表示该词项的重要性。TF-IDF 编码中的重要性数值由两个相对简单的计数统计量 TF 和 IDF 相乘得出。词频 TF 表示某个单词在给定文档i中出现的次数，通常会进行归一化，以适应不同的文档长度。相较之下，逆文档频率 IDF 表示某个词项在给定文档集合中的“信息量”。其背后的逻辑是，如果某个词项几乎出现在所有文档中，那么它对于区分不同文档所携带的信息就很少。

下一个问题是如何表示用户u的兴趣画像，也就是如何实现 *ContentBasedProfile*函数。一个常见目标是采用一种表示方式，使其能够轻松地与各个物品的内容表示进行匹配。对于文本文件，可以先用 TF-IDF 对用户过去喜欢的所有物品进行编码。然后，*ContentBasedProfile* 可以定义为这些受喜爱物品的向量的平均值。最后，既然用户画像和物品已经以兼容的方式表示，*score* 函数就可以通过余弦相似度函数来实现。

当然，不同功能也可以有各种替代实现方式。可选方案从更简单的方法（例如，在电影推荐系统中统计演员等重叠特征），到更复杂的嵌入表示不等，后者旨在更好地捕捉文档的语义。[7] 当然，也可以采用监督式方法，利用基于物品内容信息的输入特征，并使用任意合适的浅层或深度模型来学习预测函数。此外，在过去几十年中，人们还考虑了文档本身之外的各种补充信息，以更好地理解不同物品之间的相似性或关系。除了物品元数据之外，人们还将有关物品的各种形式的外部 信息（例如来自维基百科的信息）纳入基于内容的（或语义感知的）推荐系统，详见 [Musto 等人 2022]的深入讨论。

[7]虽然 TF-IDF编码在某些应用中往往能取得不错的结果，但这种简单的（“浅层”）编码有一个明显的局限：它没有赋予词语任何含义，也没有考虑词项在文档中的位置。

在应用领域方面，当可用的协同信息不足时，人们常常会使用纯基于内容的系统。一个典型的应用场景是新闻推荐，因为我们必须处理源源不断的新物品；产业界的案例研究见 [Kirshenbaum 等人 2012]。不过，基于内容的系统也已应用于电影或移动应用等其他领域 [Bambini 等人，2011，Jannach and Hegelich 2009], ，在这些领域中，用户偏好可能相对稳定，例如体现在偏好的类型上。基于内容的方法的一种特殊应用是“相似物品”推荐，其推荐依据不是用户档案，而是某个特定物品。相似物品推荐在视频流媒体平台上很常见，例如推荐与用户刚看过的电影相似的电影，参见 [Trattner 和 Jannach，2019，Yao 和 Harper 2018]。

1.3.6.2 混合推荐系统

纯基于内容的系统可能存在一定的局限性。这类系统的一个固有特征是，它们在设计上会推荐“更多相同的内容”，因此限制了帮助用户发现新类型内容的能力。此外，基于内容的系统可能会推荐过于小众的内容。例如，一个只考虑类型或剧情简介的电影推荐系统，可能会忽略电影的重要质量因素。因此，它可能会返回一些在内容上与用户过去喜欢的电影相关、但最终用户并不会喜欢的电影。

解决此类问题的一种常见方案是构建混合系统，即不仅依赖单一范式（例如协同过滤），而是结合不同方法的系统。这样可以避免单个方法的缺点，同时结合其优势。在我们的基于内容的电影推荐示例中，可以在返回推荐之前应用一些质量过滤器，例如移除所有社区评分较低的电影。

在文献中，人们已经确定了构建混合推荐系统的多种方式。在 [Burke 2002], Burke 确定了其中七种方式，随后在 [Jannach et al. 2010]中将其归纳为以下三大类别：

- 单体式：在这种设计中，不同推荐策略的各个方面被实现在同一个算法中。这种方法的一种非常常见的实现方式，是设计一个同时包含协同信号和基于内容的特征（“辅助信息”）的机器学习模型。

- 并行式：在并行式设计中，首先独立确定两个或更多算法的结果，然后以某种方式将其合并。比如，可以在用户界面层面进行合并，向用户展示一个同时包含个性化推荐和最近热门物品的列表。或者，也可以采用某种加权方法，让每件物品从每个推荐器处获得一个分数。

这种加权方法的极端情况是切换式混合推荐：根据具体情况，推荐始终取自某一个特定算法。[8]

- 流水线式：在这种设计中，一个算法将另一个算法的输出作为输入。实际应用中，例如，可以先根据流行度检索出若干推荐候选项，然后仅根据假定的用户偏好，以个性化的方式对这些物品进行排序。[9]

1.3.6.3 利用侧信息的协同过滤

最常见的混合方法可能是利用各种形式的侧信息来增强协同过滤的能力。不过，这类侧信息并不限于基于内容的过滤方法中的物品相关信息。Wu et al. [Wu et al. 2022] 将侧信息归纳为以下几类信息，推荐系统可以在观察到的用户—物品交互数据之外使用这些信息。图1.1展示了不同形式的信息。

- 用户数据：这些是用户的静态或缓慢变化的特征，例如年龄、性别或国籍。历史上，利用这类信息的系统被称为人口统计学推荐系统。此外，研究人员还探索了在推荐过程中考虑人格特质的方法，参见 [Dhelim et al. 2022]。

- 物品数据：这些是与特定物品相关的特征。文献中探讨了各种领域特定的物品相关特征，包括类别等预先结构化的信息、物品文本描述、评论或图像等非结构化信息，以及以标签形式呈现的由学界提供的半结构化信息。

- 上下文数据：文献中探讨了各种变化相对频繁的上下文信息，最重要的包括用户的地理位置、一天中的时间、天气状况或终端用户设备，参见 [Adomavicius et al. 2022]。

就用户—物品交互数据而言，过去的研究历来只关注评分这一种交互形式。近年来，基于隐式反馈信号进行推荐已成为主流，综述见 [Jannach et al. 2018] 。在大多数已发表的研究中，通常假设只存在一种隐式反馈，例如，表明用户是否与某件物品发生过交互。实际上，可用的交互类型不止一种，也都可以加以利用。在电子商务商店中，可用的信号可能包括物品浏览、加入购物车事件、购买行为、类别导航事件、搜索行为，甚至

[8]例如，在 [Jannach and Hegelich 2009], 提出的移动游戏推荐系统中，在用户购买了至少一定数量的物品之前，系统采用基于内容的技术。[9]这种分阶段的方法在拥有庞大项目目录的工业应用中很常见，例如 YouTube [Covington et al. 2016]。

之后的物品退货等。目前，关于利用多种此类交互数据的研究还比较有限，这可能是因为交互事件具有应用场景特定的性质。此外，在现实世界的交互日志中，通常会记录交互发生的时间。有了这些信息，就可以实现时间感知推荐系统[Campos et al. 2014] 以及序列感知方法，另见第 2.1 节。

最后，如上所述，文献中存在多种旨在将外生信息纳入推荐过程的方法。此类信息（包括“世界知识”）的例子有维基百科文章、公开可访问的领域本体和知识图谱，或可通过关联开放数据端点访问或查询的数据，相关讨论见 [Musto et al. 2022] 。除了这类主要与物品相关的信息外，文献也探索了与个体用户相关的外生信息，例如用户的社交网络、信任关系或用户撰写的评论，参见 [Chen et al. 2015, Dong et al. 2022, He and Chu 2010]。

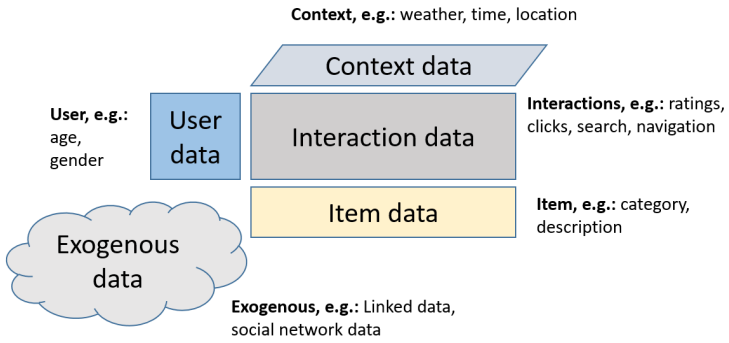


图1.1 数据类型（据 [Wu et al. 2022]扩展）。

关于构建“信息丰富型”协同过滤系统的技术方法，近期综述可见 [Wu et al. 2022] 和 [Sun et al. 2019c]。尤其是后者概述了各种深度学习模型如何纳入不同类型的侧信息，包括扁平特征、层次特征、知识图谱、图像特征等。

1.3.7 讨论

本节回顾了若干重要的推荐算法方法。需要强调的是，这仅是一个高层次的概述，读者还可以在文献中找到许多其他算法及其变体。例如，Rendle [2010, 2012] 提出了分解机框架，将包括矩阵分解在内的多种因子模型加以推广。与神经网络方法类似，分解机能够灵活地利用

用户和物品侧信息，在该框架下实现一体化的混合推荐方法。大约在同一时期，Ning 和 Karypis [2011] 在线性模型的范畴内提出了 SLIM，这是一种对物品-物品协同过滤的推广；从某种意义上说，它通过求解类似于第 1.3.3.2 节矩阵分解中公式 1.3 的正则化优化问题来学习相似度矩阵。由于在经验上效果良好，该算法已成为推荐系统文献中的常用基线。后来，Steck [2019] 推导出一种闭式解，用于计算与 SLIM 非常相似的模型。该方法省去了代价高昂的迭代优化过程，因此效率大幅提升。

其他算法方面的发展则针对特定的推荐场景，例如序列推荐 [Quadrana 等人 2018],（将在第 2.1 节讨论），以及／或者对推荐问题的重新审视，例如强化学习 [Li 等人，2016，Xin 等人，2022a],（我们将在第2.2节探讨）。

总体而言，尽管协同过滤方法尤其在实践中取得了成功，推荐问题仍远未得到解决：我们建议读者回顾自己作为推荐系统用户的亲身体验，并形成自己的看法。要让推荐体验达到现代搜索引擎的效果水平，涉及许多因素。严格来说，这种比较并不完全可行——这样的比较并不公平，因为搜索系统会收到用户明确表达的意图：查询。无论如何，与搜索竞争都会歪曲推荐的宗旨：推荐能在搜索无法解决或不应依赖搜索的领域创造价值。尽管如此，当前推荐技术仍有很大的改进空间，要实现改进，就需要从不同角度着手，正如本文后续内容将要讨论的那样。开发更好的算法无疑是其中一个方面，也是该领域关注和投入的一大重点。

1.4 推荐系统评估

如上一节所述，部署推荐系统时有许多选择，包括大量算法、算法变体及特定配置。要做出明智的选择，就需要合适且可靠的评估方法。因此，设计和选择此类可靠的评估方法，是推荐技术开发与研究的核心环节。同时，这些评估方法也推动着该领域的算法演进——作为适应度函数，它们塑造着最先进水平。

评估通常涉及对两个或多个推荐系统或其变体进行比较。在生产环境中比较系统最直接的方法是在线 A/B 测试，例如参见 [Garcin 等， 2014]。不过，在线实验是一种有限的资源，既耗费时间，也可能因用户接触到次优的系统变体而导致用户体验下降。因此，离线评估被用作一种补充手段，用于筛选出哪些变体和变更提案值得进入成本更高的

在线测试 [Gruson 等，2019，Gunawardana 等， 2022]。另一方面，离线评估能够快速、安全地进行超参数探索，并且仍是学术研究中广泛可用的主要实证手段。

在接下来的章节中，我们将概述推荐系统评估的主要方法路线。我们会简要介绍在线方法（这一视角将在后续第 2.3 节中进一步阐述），然后重点讨论离线评估。第 2.1.3 节还将讨论基于会话的推荐情境下更具体的评估方法。

1.4.1 在线评估

我们在 Spotify、Netflix、亚马逊、YouTube、Booking、Twitter、Facebook 等平台上体验到的推荐技术，都经历了一个筛选流程：从算法构想的初步形成（可能是在学术研究背景下），到最终的生产系统 [Amatriain 和 Basilico 2015]。这一流程结合了离线和在线实验，在其中，新构想会相互比较，与既有基线比较，最终还会与在线应用中当前运行的推荐算法进行比较。总的来说，鉴于实时测试的成本和带宽限制，离线实验通常先于在线评估进行 [Gruson 等人 2019]。验证新构想的最终测试，是将其与当前版本一同部署到生产平台，让一部分用户流量使用新版本，并将其表现与现有系统进行比较——胜出者将接替现有版本，或继续作为运行系统的版本。

这被称为 A/B 测试，因为会比较两个系统 A 和 B，其中 A 通常指当前系统（借用临床试验术语，常称为控制），而 B 则是挑战组（试验组）。在实践中，人们通常还会运行所谓的多变量测试，同时比较两个以上的系统。正如本节开头提到的那些热门应用一样，目前正在同时运行数百项 A/B 测试 [Amatriain 和 Basilico 2015]。A/B 测试的一个真正特征是，用户不知道自己使用的是哪个版本的系统，就像药物研究中的受试者不知道自己接受的是正在试验的实际治疗（系统 B）还是安慰剂（系统 A）一样。这样做旨在通过避免在实验中引入混杂偏差，最大限度地提高比较的客观性。

除了在尽可能贴近真实的条件下测试想法，A/B 测试还可以衡量算法对最终业务绩效和目标的影响，而不仅仅是理论性较强、推测性较大的效果指标。影响可以通过参与度（点击量、播放次数、停留时间）、销售额（收入、利润、购买次数）、客户留存，以及企业所依赖的任何其他指标或关键绩效指标来衡量。我们还将在第 2.3 节从更广泛的角度进一步阐述推荐的商业价值，同时考虑客户和系统之外其他可能参与推荐并受到推荐影响的利益相关者。

A/B 测试通常会持续进行，直到发现系统 A 和系统 B 之间存在统计显著差异，或在没有明确结论的情况下达到可承受的时间上限。A/B 测试通常持续数周——这是收集足够证据、以便就系统变更做出数据驱动决策所需的时间。测试时长以及让客户接触未经测试的新变更所涉及的风险，限制了在生产平台上同时开展的实验数量。因此，离线评估会用于筛选出少量变更提案，再将其提交在线评估——我们将在下一节详细讨论离线实验方法。

在这方面，一个尚待解决的挑战是，离线与在线比较的结果之间常常呈现出较弱的相关性，甚至毫无相关性 [Garcin 等人，2014，Gomez-Uribe 和 Hunt，2015，Jannach 和 Jugovac，2019，Kouki 等人 2020]。尽管这仍是该领域一个重大的开放性研究问题 [Gilotte 等人，2018，Jannach 和 Bauer，2020，Rossetti 等人 2016], ，在产业界，离线评估通常被用作在线测试前的筛选标准；而在学术研究中，它则是实证观测的主要手段，远胜于其他方式。

除了 A/B测试，还可以根据特定目的设计其他形式的在线实验和用户研究 [Knijnenburg 和 Willemsen 2015],，且不一定需要涉及生产系统。我们将在第2.3.1节讨论这类有人在回路的研究。

1.4.2 离线评估

离线实验可以看作对系统与用户交互过程的模拟，其中会测量用于评估系统有效性的不同代理指标——即离线指标 [Castells 和 Moffat 2022]。最有信息价值的离线实验，是最能模拟和体现生产环境的实验。在产业界的开发过程中，可能需要非常具体的生产环境。因此，在一般研究中，通常会进行一些抽象。

在线评估与离线评估的区别可以简明地定义如下：在线评估会在推荐之后 交付给用户后，从这些用户那里获取用户反馈数据（用于指标计算）；而在离线评估中，测试数据是在生成推荐之前 收集的——事实上，用户从未收到过所评估的推荐。

乍看之下，借助机器学习和信息检索等领域中广泛采用且成熟完善的方法论与实验设计原则，为评估推荐算法设置基础离线实验似乎是件直截了当的事情。然而，推荐任务有其自身的特殊性，因而存在相当程度的隐性复杂性和陷阱，比人们想象中更容易产生不一致的评估结果 [Ca˜namares 等人 2020]。我们建议实验人员仔细考虑、理解并报告实验中所作的详细设计选择及其影响。下面几节将对其中一些内容进行概述。

1.4.3 离线数据

离线评估通常会区分训练数据和测试数据，二者应严格互不重叠 [Guy 2022]。训练数据作为输入提供给待评估的推荐系统，而测试数据对系统隐藏，用于计算返回的推荐的指标。离线实验中的训练数据可以类似于生产系统可能拥有的数据（越相似越好），而测试数据则用于在实验中模拟用户对所提供推荐的反馈。用于推荐的训练数据和测试数据通常都将用户-物品交互作为其主要组成部分之一——也就是说，它们可以被视为评分矩阵中互不重叠的子集。不过，训练数据还可以包含特定推荐算法可能使用的其他辅助信息。

训练数据和测试数据可以通过许多不同的方式获取。构建训练集的一种好方法，是导出真实推荐系统在某个时间点所使用输入数据的一个子集。多年来，许多此类数据集已公开发布，其中最著名的是非商业系统 MovieLens 的数据集[10]。如今，研究人员还使用各种其他数据集，其中包含物品评分、购买信息或收听事件，另见 [Harper 和 Konstan 2015][11]。最近还发布了若干数据集，其中不包含用户-物品交互矩阵，而是记录交互的序列日志，可用于评估基于会话的推荐算法，参见第 2.1 节。遗憾的是，对于许多数据集，我们并不清楚它们是在何种具体情况下收集的。例如，如果某个电子商务或音乐流媒体平台在数据收集时已经部署了推荐器，那么日志中观察到的内容可能在一定程度上受到现有推荐算法的偏差影响。无论如何，对于这样一个包含日志交互的数据集，获取测试数据的一种简单方法是从训练集中进行子采样。这通常称为对数据进行划分，可以通过不同方式进行，我们将在下一小节讨论这些方式。

通常为推荐系统实验收集的数据会呈现严重的抽样偏差，这些偏差源于收集数据所依托的运行系统（其算法、用户界面和业务规则）、外部偏差（营销、时尚、社会影响等）以及用户参与中的行为偏差。这引发了评估方面的问题，我们将在第2.2节稍后讨论这些问题。不过，测试数据也可以从独立于训练集的其他来源获取。例如，Yahoo! R3 数据集 [Marlin 和 Zemel 2009] 包含一个测试集，其中的评分是从用户对音乐的评分中随机抽样得到的——这意味着系统要求用户对为其均匀抽样的音乐进行评分——而训练集则通过用户与音乐之间的自由用户交互收集。Coat 数据集 [Swaminathan 等人2017] 以类似方式在服装领域构建。CM100k 数据集 [Ca˜namares

[10] https://www.movielens.org[11]还可以在 https://github.com/RUCAIBox/RecSysDatasets 找到其他数据集。

和 Castells 2018] 完全随机地收集了用户对音乐的评分，并提供了一个关于用户对音乐熟悉度的补充标签，该标签被认为可以作为非均匀分布输入训练数据的代理指标。

1.4.3.1 数据划分

当用于离线评估的数据不包含单独的测试子集时，通常会从现有数据集中子采样得到测试子集。抽样（称为划分）可以通过不同方式进行。划分过程的首个参数是划分比例，即训练集与测试集数据的比例，通常以$[0,1]$中的百分比或比率表示。鉴于训练推荐模型时通常面临数据稀疏性问题，训练集通常会分配较多数据（例如 $\geq 80\%$），而测试集则较少。

抽样过程中的第二个选择维度是随机抽样还是时间划分。对于选定的时间点，时间划分会将该时间点之前产生的所有数据放入训练子集，其余数据放入测试子集 [Koren 2009]。有时会根据有意义的时间单位（例如数周、数月等时长的数据）来选择时间点，也可以选择一个能得到特定划分比例的时间点（即训练数据与测试数据量之间的比例）。研究人员也探索了更复杂的变体，例如将数据划分为多个时间窗口 [Lathia 2010]。时间划分通常能提升评估方法的可靠性，因为它更能体现真实的推荐任务：根据过去观察到的用户行为，预测未来（或当前）的有用性。此外，它还能实现更清晰的数据分离，因为混合过去和未来的记录可被视为一种数据泄漏。就此而言，采用统一的时间划分点比针对不同用户采用不同划分点（例如使用留一法（最后一次交互））更加清晰。否则，我们可能会根据前瞻信息的泄漏，预测用户将来会喜欢某件物品；例如，根据划分时间点较晚的用户的训练记录得知该物品届时会变得流行或时髦，而实际上，在该用户的训练期内，这件物品甚至可能尚未被创造出来。

不过，时间划分并非总是可行，也未必是严格必需的。例如，可能没有有意义的数据时间戳，或者用户需求可能相对稳定，又或者实验者可能希望聚焦于某个特定问题，在这种情况下，时间并非优先考虑的变量。常见的替代方法是根据划分比例，随机抽样测试数据和训练数据。与时间划分相比，随机抽样更加灵活，例如可以进行n折交叉验证，其中$n = 5$是一个典型的取值（例如$n = 10$也很常见）。

一种自然的随机划分实现方式是对数据记录进行独立同分布的伯努利抽样（抛硬币）B(1，p)，其中$p =$为划分比例。文献中还记载了其他实现方式：以均匀随机且不放回的方式抽取确切数量的数据记录，有时会针对每个用户或每个物品分别抽取。其他作者则尝试为每个物品或每位用户抽取相同数量（而非相同比例）的测试数据，以此来

减少评估偏差 [Bellog´ın 等人 2017]。针对特定问题的研究也可以采用正交变化的划分流程，通过将关注的物品或用户放入测试集，来模拟冷启动或长尾等特定情形 [Hurley 和 Zhang 2011]。

1.4.3.2 候选物品采样

离线实验设计中一个不太显眼的选项，是为每位目标用户选择一组候选物品（有时称为目标物品 [Ca˜namares and Castells 2020, Sarwar et al.2001]），供待评估的系统进行排序。一种自然的做法是选择目录中的所有物品。不过，当推荐被视为矩阵补全问题时，情况并非如此：已知的矩阵单元格并不需要补全。换言之，如果推荐的目标包括帮助用户发现新物品（大多数文献都是如此），那么对于训练记录中已有某物品的用户，不会将该物品纳入推荐结果。

实验人员也可以考虑将候选物品限制在更小的集合中。Koren[2008] 最先提出为每位用户设定固定数量的目标物品。这一想法逐渐流行起来，如今在研究文献中仍很常见 [Ca˜namares and Castells 2020, Krichene and Rendle 2020]。使用较小的候选集合时，有些作者将这种做法称为计算抽样指标 [Kricheneand Rendle 2020]。一种极端做法通常称为压缩排名 [Buckley and Voorhees. 2004],，即只将带有测试评分的物品纳入候选集合——当使用误差指标时（见上文第 1.4.4.1 节），采用的就是这种做法，因为没有可用评分时无法计算评分误差。

近期研究表明，候选采样会对离线评估结果产生重大影响，因此值得深入了解 [Ca˜namares 和 Castells 2020年、Krichene 和 Rendle 2020年、Steck 2013]。作者一致指出，当候选集中纳入所有测试集中未评分的物品或完全不纳入这类物品时，系统比较结果会出现分歧。Ca˜namares 和 Castells [2020]进一步表明，在这种设置下，两个极端各有其缺点，并建议在压缩排序与完整排名之间选择某个点，以优化离线评估的信息量。

1.4.4 推荐任务与指标

实际上，每个推荐系统的设计都有特定目的，以便为用户和组织创造价值，另见第 2.3 节。根据目的，系统中需要实现不同的计算任务 [Jannach and Adomavicius 2016]。在推荐系统评估领域的开创性研究中，Herlocker 等人 [2004]指出了诸如“找到优质物品”“上下文中的注释”（预测评分）或“推荐序列”等任务。在设计新算法时，核心问题是了解这一新算法在多大程度上能够有效完成这些任务。在离线实验中，这种评估借助计算指标完成；这些指标充当替代指标，用于衡量系统在其使用情境中的效果。

传统上，推荐的首要总体目标被认为是准确把握用户需求。因此，为评估这一点而设计的指标通常统称为准确率 指标。如今，推荐准确率被视为排序质量的同义词，因此，指标和评估协议借鉴并改编自信息检索领域 [Bellog´ín 等人 2017],的相关工作，正如我们将在第1.4.4.2节中讨论的那样。不过，出于历史原因，我们将先讨论早期基于评分预测的观点。

1.4.4.1 评分预测

如第1.2节所述，推荐任务最初被理解为评分预测问题，也就是说，这是一个回归任务，需要根据示例学习一个函数 $f: U \times I \to R$。因此，使用平均绝对误差（MAE）和均方根误差（RMSE）等误差指标来评估推荐效果，似乎顺理成章。误差是在评分矩阵的测试数据上计算的：

$$\text{MAE} = \frac{1}{|T|} \sum_{(u,i) \in T} |f(u,i) - r(u,i)| \qquad \text{RMSE} = \sqrt{\frac{1}{|T|} \sum_{(u,i) \in T} (f(u,i) - r(u,i))^2}$$

其中，$T \subset U \times I$ 表示测试数据记录的子集；这些指标的值越低，推荐系统的效果就越好。

在近二十年间，推荐系统文献一直采用这类指标，它们也成为推动该领域发展的重要举措——例如 Netflix 奖 [Liu et al. 2007]——的基础。尽管如此，学界已不再局限于评分预测，如今倾向于将推荐视为并按排序任务来处理，这也符合产业界占主导地位的商业模式 [Cremonesi et al. 2010]。另一方面，许多推荐算法会为没有有意义评分解释的用户-物品对输出分数，但这些分数非常适合用于筛选出要推荐的高分物品列表；误差指标并不适用于这类分数。

1.4.4.2 排序质量：将推荐视为信息检索任务

在 2000年代，学界越来越认同这样一种观点：提供推荐与提供搜索结果有许多共通之处 [Herlocker 等人 2004]。推荐系统和搜索系统都帮助用户从大型集合中获取相关信息或产品。两者的一个主要区别在于，推荐任务没有显式查询——不过，仍可将其理解为需要满足用户需求的问题，即使用户没有明确表达这种需求。事实上，在许多应用中，搜索与推荐往往协同工作、相辅相成。早在 [Belkin 和 Croft 1992] 之前，人们就已认识到这一观点，但它似乎直到 2010年代才在整个学界逐渐普及，并在一定程度上推动了评估范式的转变 [Bellog´ín 等人，2017年；Cremonesi 等人 2010]。

认识到这一点后，推荐效果的关键就在于返回结果的排序：有效的推荐应尽可能让更多相关项尽早出现在排序中，并尽可能减少排在前列的不相关项。排序决定了用户浏览推荐内容时会看到哪些物品，以及何时看到。将系统分数理解为准确的评分预测，已无关紧要。排序这一概念还可以自然地推广到其他不一定呈线性排列的推荐展示形式（例如“货架式”矩阵）；在这类形式中，屏幕布局中的某些区域——相当于排序靠前的位置——比其他区域更有可能被用户查看。

排序评估自然促使研究人员借鉴信息检索领域的概念和方法。半个世纪以来，该领域一直在研究和发展离线评估流程与指标。然而，这种改编并非简单直接，其中涉及一些需要谨慎处理的细微之处。我们将通过讨论信息检索领域离线评估实践中的主要要素，来探讨这种改编[Sanderson 2010]。

集合。 物品集合 I 可以视为信息检索文献中所有“文档”所构成集合的对应物。物品集合通常也称为“项目目录”；事实上，在大多数推荐系统应用中，它们往往是搜索与推荐互补功能所共用的检索空间。

查询与信息需求。 搜索和推荐都源于用户一侧的需求，而系统旨在帮助满足这些需求。用户会主动通过显式查询向搜索引擎描述自己的需求；相比之下，推荐系统则从观察到的用户活动以及用户与物品的交互中推断用户需求。因此，与搜索引擎相比，推荐系统中的信息需求表征相当模糊且不完整，也因此需要采用不同的检索技术。搜索查询也可能具有不同程度的模糊性；在检索任务中，推荐可以视为查询具体性连续谱的一个端点，此时查询为空。

相关性。 相关性的概念是信息检索的核心，在推荐中同样有意义，其含义也完全相同：如果用户喜欢、享受、对某个推荐物品感到满意或感兴趣等，那么该物品就与用户相关。可以将相关性视为推荐为用户带来任何价值的必要条件。如果推荐的物品与用户无关，用户就会忽略它，也不会从中获得任何收益。

不过，应当指出，在离线实验中处理相关性的方式存在一个重大差异：面对查询时，为了合理简化，通常假定相关性是客观的，也就是说不因用户而异；但在推荐中，这一假设并不合理，因为相关性被认为高度依赖用户，这通常也是推荐任务的内在特征。这一点对离线评估中的相关性判断有重要影响：将物品标注为相关或不相关的工作委托给评估人员，

让他们代表用户作出判断，这种做法实在过于牵强，无法对个性化推荐进行任何有意义的评估。因此，离线评估无法像搜索评估那样与实际最终用户分离。

在离线推荐实验中，从目标用户获取的测试数据在信息检索评估中起到相关性判断的作用。与信息检索中的判断汇聚相比 [Sanderson2010], ，当数据是在真实环境中收集，和／或通常情况下由推荐系统介入收集时，测试数据可能导致评估中出现极其显著的选择偏差。测试数据在物品间的分布极不均衡，呈现出所谓的流行度偏差，可能会严重扭曲评估结果 [Bellog´ın et al. 2017, Ca˜namares and Castells 2018, Jannach 等人，2015b], ，我们将在第2.2节中进一步讨论这一问题。

指标。一旦定义并获得了与相关性判断等效的数据，就可以将任何信息检索指标应用于推荐系统的输出：精确率、召回率、平均准确率均值（MAP）、平均倒数排名（MRR）和归一化折损累计增益（nDCG）常用于评估推荐结果的排序质量。由于精确率和召回率是不考虑排序的指标，通常会在排名前n 个推荐物品的子集上进行测量——即排序截断点——记作 P@n 和 Recall@n。对于 nDCG、MAP 和 MRR 等考虑排序的指标，也可以设置截断点，以便进一步聚焦于排序靠前的结果。

ROC曲线下面积（AUC） [Bamber 1975] 也可以被视为一种排序指标，因为它从排序的角度展现了推荐的相关物品与不相关物品之间的权衡。AUC 通常被用作分类器评估的指标，但精确率和召回率也是如此：它们将推荐系统视为将物品分为相关类和不相关类的二元分类器。将测量指标聚焦于排名靠前的n 个位置，就能实现这一点，使它们转变为信息丰富的排序指标。

与搜索实验相比，推荐相关性判断中的偏差和用户依赖性加剧了离线推荐实验中待评估的排序顶部区域的相关性标签稀疏问题。正如前文第 1.4.3.2 节所述，决定如何处理缺失的相关性判断，成为引入信息检索排序指标时的一个关键问题。另一个挑战是缺失数据并非随机产生，从而导致评估结果出现偏差。我们将在第2.2节进一步讨论这一问题。

1.4.5 超越准确率

长期以来，准确率一直是算法评估的主要视角，但后来人们逐渐发现，这并不能全面说明推荐为何能为消费者和提供者带来实用价值和利润 [Ge 等，2010年，McNee 等 2006]。例如，披头士的粉丝很可能会喜欢《 Yesterday 》这首歌 —— 从信息检索的角度看，它高度相关——但把它作为推荐内容又有多大用处呢？这首歌是全球最广为人知的音乐作品之一，对披头士的粉丝来说更是如此。用户想听的时候，完全可以自己搜索。这首歌作为一种便利型推荐或许有一定意义

在某些场景中，例如驾车出行时使用自动播放列表，但与搜索和浏览等方式相比，这类推荐所增加的价值显然有限；相比之下，发现新的音乐、新的乐队，或发现用户此前未曾意识到、如今却能欣然享受的新品味，可能更有价值。

同样，当流媒体平台列出用户可能喜欢的电影和剧集时，纳入不同类型和不同导演的电影或许是明智之举，这样能更好地体现人们兴趣的多样性，以及他们心情的变化。虽然人们的品味可能相对稳定，但我们可能今天想看喜剧，明天又想看纪录片，而这些变化很难预测。与其冒险孤注一掷，只推荐清一色的喜剧或纪录片，不如提供多样化的潜在心头好，这似乎是更明智的做法。此外，通过让推荐更加多样化，系统也有机会探索用户尚未显露的兴趣并加以了解，从而在未来持续改进。

新颖性与多样性只是准确率之外、能够最大化推荐价值的两个目标维度，也是本文的关注重点。区分新颖性与多样性很有意义：二者密切相关，却代表不同视角，这一区别也体现在下文的总结中。关于推荐所带来的价值、用户满意度、业务绩效，以及对相关利益相关者产生的广泛影响等直接关键衡量指标的更广泛讨论，见第 2.3 节。推荐系统有多种动机来提升新颖性与多样性，也可以从不同角度理解这些一般性概念，并将其转化为相应的指标和算法，以提升和衡量新颖性与多样性，接下来我们将对此展开讨论。关于推荐系统中的新颖性与多样性的更全面综述，请参阅 [Castells 等人 2022]。

1.4.5.1 新颖性

从抽象层面而言，推荐新颖性可定义为推荐物品与某个参照上下文之间的差异 [Vargas 和 Castells 2011]。上下文通常可以是接受推荐者的过往经历；或者，考虑到系统对该用户的了解稀疏且不完整，也可以是系统中所有用户的汇总经历。在后一种情况下，新颖性等同于稀有性：如果某个推荐物品属于交互频率分布的长尾 部分，那么它就是新颖的 [Celma 和 Herrera2008]。个性化新颖性通常也称为意外性 [Adamopoulos 和 Tuzhilin2014], ，反映的是某项推荐与观察到的目标用户过去交互过的物品之间的差异。长尾新颖性和意外性可以通过不同方式量化，接下来我们将对此进行总结。

长尾新颖性。 例如，可以将这一维度衡量为一个单调递减函数 ϕ 的平均值，该函数以过去对每个推荐物品的参与度作为自变量L_i

:

$$\mathrm{LT}(R) = \frac{1}{|R|}\sum_{i \in R} \phi(|L_i|)$$

其中，R 是返回的推荐，L_i 表示涉及物品i的已观察到的交互集合。文献中对 ϕ 函数的定义包括：涉及该物品的已观察到的交互占比的补数，或该占比的负对数 [Vargas 和 Castells 2011]:

$$\phi(x) = 1 - \frac{x}{|L|} \qquad\qquad \phi(x) = -\log\frac{x}{|L|} = \log\frac{|L|}{x}$$

其中，L 表示所有已观察到的交互集合。根据第一种定义，新颖性可以理解为：随机用户过去从未与R中的推荐物品交互过的概率。忽略重复交互时（即在上述公式中，每个用户-物品对最多计数一次），第二种定义等同于信息检索模型中文档检索所定义的逆文档频率；此处用户相当于文档，因此有时也称为逆用户频率 [Breese 等人 1998]。以这种方式衡量时，新颖性可视为冷启动（数据缺乏）或不受欢迎程度的一种表现。

意外性。 这一概念通常量化为：

$$\mathrm{Unex}(R) = d(R, E_u)$$

其中，d 表示集合差异度算子，$E_u \in I$ 是一个参考物品集合，用于表示每个用户$u \in U$的“预期内容”。例如，预期集合可以是目标用户交互过的物品、某个被认为较为常规的算法所给出的推荐，或任何其他可代表不令人意外体验的替代指标 [Adamopoulos 和 Tuzhilin2014]。集合差异度可以采用任何常见度量，例如杰卡德、豪斯多夫、平均逐物品距离，或其他有意义的扩展，以量化两个集合之间的差异。使用成对物品距离时，基于特征的差异度（例如，通过特征向量/集合之间的余弦距离或杰卡德距离来衡量）通常是不错的选择。其他基于集合的差异度算子也可以具有概率解释，例如，$d(R,E_u) = |R \setminus E_u|/|R|$ 可以理解为推荐物品不在预期之中的概率。

意外发现。 在新颖性的范畴内，另一个重要概念是意外发现。尽管文献中的定义略有不同，但主流观点认为，如果一项推荐既令人惊喜（即新颖）又有价值，那么它就是意外发现 [Chen et al. 2019]。如果在此意义上将价值等同于相关性，衡量意外发现的一种直接方法是：仅对与目标用户相关的推荐物品（根据现有测试数据判断）计算上述新颖性度量指标，并忽略其余物品。

其他概念。 上述新颖性概念或许是文献中最常见的，但并非面面俱到 [Castells et al. 2022]。例如，新鲜度 （推荐物品创建于何时）本身往往是一项重要属性，通常也是与上述概念相关联的一种信号。Lathia 和 Amatriain [Lathia et al. 2010]探讨了更细致的时间新颖性与多样性概念，这些概念涉及过往推荐的历史，广义而言，旨在衡量推荐系统在多大程度上重复自身。

1.4.5.2 多样性

与新颖性概念相关但又有所不同，多样性通常指推荐内容所涵盖的多样程度。例如，Enya 的《Orinoco Flow》和 AC/DC 的《Highway to Hell》很难被视为新颖之作（毕竟谁都曾听过这些音乐），但将它们放在一起，会构成多样性很高的音乐推荐，因为两者截然不同。同样，两首非常相似的稀有音乐作品虽然可能极具新颖性，却不具备多样性。

列表内不相似度。 衡量内部多样性的一种常见且直接的方法，是计算所推荐物品之间成对差异度的平均值 [Smyth 和 McClave，2001，Ziegler 等人 2005]:[12]

$$\mathrm{Div}(R) = \frac{2}{|R|(|R|-1)} \sum_{i \in R} \sum_{j \in R} d(i,j)$$

其中 $d : I^2 \to \mathbb{R}$ 是一个通常基于物品特征的差异度函数。根据这一定义，多样性可视为意外性的一种特殊情况，其中预期集合就是推荐本身：$E_u \leftarrow R$。

基于方面的多样性。 或者，也可以根据某个特别有意义的特征来定义推荐列表中的多样性，这种特征被称为方面空间，例如电影类型、音乐风格、国籍等。所谓的基于方面的指标，如方面召回率 [Zhai 等人 2003] 提出的指标，或意图感知指标框架 [Agrawal 等人 2009] （在搜索系统的上下文中定义），都可以直接用于评估推荐多样性[Vargas 等人，2011年，Wasilewski 和 Hurley 2016]。

覆盖率。推荐中的另一个重要多样性维度，是从用户总体层面考察推荐内容在物品空间中的分布。这一多样性视角通常称为覆盖率、总体多样性或销售多样性。用于衡量这一概念的指标（无论是否有意）借鉴自生态学领域 [Patil andTaillie 1982]，，从非常简单的公式（至少向一位用户推荐过的物品数量

[12]Ziegler 等人 [2005] 和其他作者将列表内相似性而非不相似性定义为需要最小化而非最大化的指标。这两种观点在某种程度上可以视为（显式或概念上）互补，正如相似性与距离可以互为补充一样。

）到更精细的度量，例如基尼指数 [Gunawardana et al. 2022]。这一多样性视角日益受到重视，因为它与公平性和机会均等问题直接相关：所有物品（及其提供者、供应商或创作者）都有机会展示给消费者。

1.4.5.3 提升新颖性与多样性

提升新颖性与多样性，可以通过最大化相关指标来实现，即将上述讨论过的指标作为优化目标。这会带来两个主要问题。第一个问题或许也是最重要的问题在于，从信息检索的意义上说，最大化新颖性或多样性与最大化相关性并不一致。事实上，这两个目标有时会相互制衡——例如，随机推荐通常在新颖性和多样性方面得分极高，而在相关性方面得分极低。该问题呈现的并非单一最优解，而是一个帕累托前沿。作为一个多目标问题，多样性与新颖性增强可以通过常见策略来实现，例如标量化、进化算法等。 [Veloso 等人 2014]。由于最先进的推荐算法不一定处于帕累托前沿，一些研究报告称，在这两个方面都取得了改进[Jannach 等人 2015a、Vargas 和 Castells 2014]。

据此，可以对多样化策略作出一个高层次的区分：内在策略与外在策略。外在方法会对初始的、以相关性为导向的推荐进行重新排序，以期改善相关性—多样性（和/或新颖性）权衡。这类方法中一种基本且典型的做法，是通过贪心最大化相关性与多样性的线性组合，对初始推荐进行重新排序。在内在方法中，相关性和多样性会作为算法的内置目标，同时在内部加以处理 [Vargas 和 Castells 2014]。例如，可以在多目标函数中同时纳入新颖性和/或多样性以及准确率，并对该函数进行最大化 [Veloso 等人 2014] ，这与前文第 1.3 节讨论的思路一致。

第二个难点在于，多样性指标并非针对每个用户单独定义后再进行汇总或平均，而是针对全部推荐项构成的集合定义——覆盖率就是前文解释过的一个例子。在这种情况下，无法针对每个用户独立优化相关性—多样性权衡，问题空间也会变得更大：各项推荐对全局多样性的贡献彼此依赖，理想情况下应同时进行优化。不过，这种方法仍可采用不同程度的贪心策略，但还需要在成本与最优性之间进行权衡 [Sanz-Cruzado 和 Castells 2018]。

近年来，在个体用户层面运作的多样化方法越来越受欢迎。早期的一项研究中，Oh 等人 [2011] 提出，应根据个体用户过去对流行度的倾向来调整推荐的新颖性。随后， [Jugovac 等人 2017],提出了一种类似但更通用的重新排序方法，该方法还支持为每位用户并行考虑多个优化目标。

如今，这类方法被称为校准推荐 [Steck 2018]。校准推荐的一个重要方面目前尚未得到深入探讨：用户对新颖性的偏好多样性可能会随时间变化，也可能因上下文而异。关于音乐领域中此类现象的分析，请参见 [Kapoor 等人 2015] 。

1.4.5.4 用户对多样性与新颖性的感知

过去十年间，人们提出了大量算法方法，通过根据相应的计算指标优化建议列表，生成更新颖、更多样的推荐。然而，与任何计算指标一样，确保这些指标确实能够有效反映用户的感知（例如用户对推荐的新颖性与多样性的感知）十分重要。此外，了解这些感知如何影响用户对系统的满意度也至关重要。

如上所述，列表内不相似度在文献中被广泛用于衡量推荐列表的多样性。在 [Ziegler 等人 2005], 的研究中，研究者借助一项用户研究[13]探讨了主题多样化对用户满意度的影响。他们尝试了不同程度的多样化，结果发现满意度达到峰值时，参与者偏好图书推荐具有一定程度的多样化。在电影推荐的上下文中， [Castagnos 等人 2013]也得出了类似的观察结果。在他们的研究中，作者向研究参与者展示了多样性程度各异的推荐列表。作者发现：（i）参与者能够察觉不同的多样性程度；（ii）多样性可能会提升用户满意度。不过，对于多样化的列表，研究发现，提供额外的解释或许有助于用户更好地将推荐与自己的偏好联系起来。 [Ekstrand 等人 2014]也报告了电影领域中多样性对用户满意度的积极影响；然而，这项研究还发现，新颖性提高到一定程度（例如，推荐了太多用户不熟悉的物品）可能会对用户满意度产生负面影响。

在本书提及的图书和电影领域研究中，所使用的相似度函数基于主题类别和电影属性，并且已在实验中证明有效。然而，其他研究中往往缺少对此类特定相似度函数的验证。因此，还需要开展更多研究，以了解物品的哪些特征决定了用户对相似性的感知，参见 [Trattner 和 Jannach，2019年、Yao 和 Harper 2018]关于电影和食品领域的研究。此外，在传统的列表内不相似度度量中，评估给定列表的整体多样性时，多样化元素的位置并不重要。实际上，有迹象表明，物品的排序可能会影响用户感知 [Ge 等人 2012]。沿着这一思路，Vargas 和 Castells [2011] 借鉴了早期基于用户模型构建的信息检索指标研究（例如， [Moffat 和 Zobel 2008]），提出了

[13]有关推荐系统用户研究的深入讨论，请参见第2.3.1节。

一种概率度量框架，可为各种新颖性与多样性指标——包括本节提及的指标——赋予排名敏感性。

不同于像[Ge 等人那样改变一个推荐列表中物品的位置， 2012], Hu 和 Pu [2011] 建议将物品组织成组。在一项同时跟踪参与者眼动的用户研究中，他们发现，“组织”界面有助于提升用户的多样性感知。最后，在一项较新的研究中，Chen 等人 [2019] 报告了一项关于电子商务平台上推荐质量感知的大规模用户研究结果。研究显示，意外发现在用户满意度中发挥着决定性作用，其重要性高于单独的多样性和新颖性。

2 近期主题

2.1 序列与基于会话的推荐

如第1.2节所述，推荐系统的传统问题形式旨在根据稳定的长期用户偏好，提供非上下文物品推荐。例如，对于媒体流媒体网站的常客，提供登录页推荐就是一种有意义的做法。然而，许多其他重要的应用场景中，我们无法假设能够获得长期偏好信息。例如，电子商务商店的访客可能尚未登录，或者是首次使用的用户。此外，在电子商务网站上，用户的兴趣和需求可能会随着购物会话的不同而变化。因此，即使我们知道用户的身份，在决定接下来向用户推荐什么时，也必须考虑用户的短期历史和当前意图。

2.1.1 问题定义与术语

传统的矩阵补全抽象难以轻松满足这些要求，因此促成了不同类型的序列感知推荐系统的开发 [Quadrana 等人 2018]。这类推荐器的主要输入不是用户-物品评分矩阵，而是一系列日志交互。这些日志条目通常通过应用程序收集，例如网络商店，而且可以有不同的类型。例如，网络商店可能会记录物品浏览、加入购物车事件或购买行为。此外，每条日志记录通常都与某个个体用户相关联。该用户可能是已登录网络商店的已知用户，也可能是仅通过 Cookie 或 IP 地址识别的匿名用户。与用户-物品评分矩阵相比，另一个区别是，交互日志可能包含多个用户-物品交互对，例如用户在购买前多次浏览某件物品。此外，日志条目还可能附带其他类型的元数据，例如时间戳。总的来说，尽管输入不同于矩阵补全问题，序列感知推荐器的输出通常是物品排序列表。图2.1展示了该问题设定的概览。

在序列感知推荐系统这一类系统中，文献中可以辨认出两个相关术语：（*i*）序列推荐和（*ii*）基于会话的推荐。序列推荐指的是更一般的下一物品推荐问题。一个典型例子是在旅游场景中，向用户推荐下一个兴趣点（POI）；或者在电

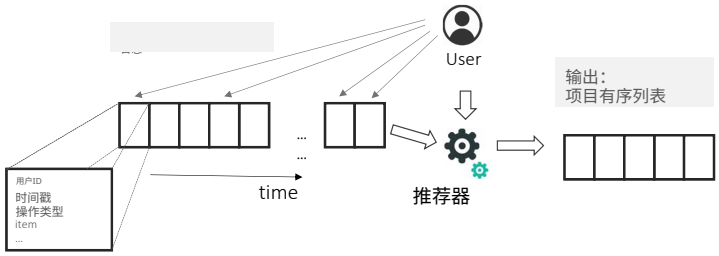


图2.1 序列推荐：概述（改编自 [Quadrana 等人 2018]）。

商场景中推荐下一个完整的购物篮。在基于会话的推荐场景中，问题同样是进行下一物品预测，但其基础假设是：（*a*）过去的交互日志按使用会话组织；（*b*）推荐是在当前会话的上下文中作出的。基于会话的方法通常还假设，用户的兴趣和需求会随着会话的变化而改变。此外，在许多实际问题场景中，当前用户完全没有长期偏好信息，这意味着推荐必须基于当前会话中观察到的一小组交互，例如少数几次物品浏览事件。

文献中，针对当前会话进行推荐时，通常会考虑两种不同的情形。一种情形是，除了会话中观察到的交互外，对当前用户一无所知。另一种情形是，还可以获得同一用户历史会话的信息。第二种情形通常称为个性化基于会话的推荐 或会话感知推荐。图 2.2 展示了（纯）基于会话的推荐问题与会话感知推荐问题之间的区别。

2.1.2 序列和基于会话的推荐算法

针对序列感知推荐问题，人们探索了各种算法方法。 [Quadrana 等人2018]指出了以下几种主要的技术方法类型：（*i*）序列学习方法（*ii*）序列感知矩阵分解（*iii*）混合方法，以及（*iv*）最近邻和其他方法。

序列学习方法。这一类别中的方法可以进一步细分为若干主要范式。

- 频繁模式挖掘技术过去常用于检测购物篮中的（意外）模式，例如以关联规则的形式。后来，这些技术被扩展到处理序列模式，其中元素的顺序很重要。

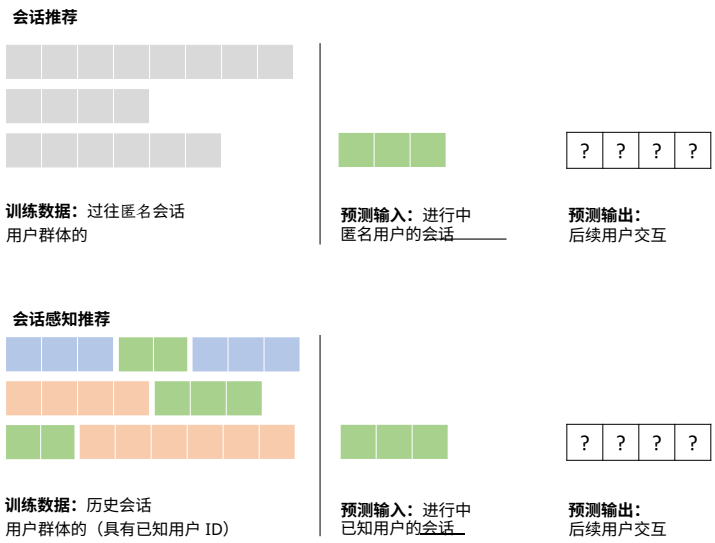


图 2.2 基于会话和会话感知的推荐：概述（改编自 [Latifi 等人2020]）。

• 序列建模方法试图超越挖掘出的模式，旨在从过去按时间排序的数据中学习模型，以预测未来事件。多年来，人们提出了各种各样的机器学习模型。文献中的主流方法基于马尔可夫模型、强化学习或深度学习，尤其是采用循环神经网络（RNNs）的形式。现代序列模型通常借鉴自然语言处理中的思路，对交互数据中的序列进行建模，参见例如 [Sun 等人，2019a]。

•分布式项目表示 是对序列信息进行编码的另一种形式。这些表示（嵌入）是将项目相关事件序列（例如，浏览）投影到较低维度的表示空间中，同时保留物品之间的转移关系。

• 滑动窗口监督学习则是一种相当具体的方法，其思路是通过在事件序列上移动滑动窗口，将下一项预测问题重新构造为监督学习问题，并从该滑动窗口中提取相应的特征值，以预测下一个事件。

序列感知矩阵分解。 文献中有一些例子，将矩阵分解方法扩展为考虑序列信息，这些信息通常是根据用户-物品评分矩阵条目中可用的时间戳推导出来的。这些方法

在概念上与时间感知 推荐系统相关 [Campos 等人 2014], 旨在识别用户偏好在较长时间跨度内的变化。

混合方法。 文献中更常见的一种方法，是将序列建模与潜因子模型的强大能力结合起来。较早采用这一思路的方法之一是 Rendle 等人提出的 因子分解个性化马尔可夫链（*FPMC*）方法 [Rendleet al. 2010]。该方法联合分解一阶马尔可夫链和传统的用户-物品评分矩阵：前者用于建模特定用户的物品转移，后者则用于生成下一物品预测。此后也提出了类似方法，例如 [He 和 McAuley 2016] 的研究。近年来，将神经网络技术（如循环神经网络、卷积神经网络或注意力机制）与矩阵分解层结合的方法越来越受欢迎，例如 [Kang 和 McAuley，2018，Tang 和 Wang 2018]。

最近邻及其他方法。 尤其是在基于会话的推荐中，一种简单而有效的方法通常是采用基于邻域的技术。在 [Jannach 和 Ludewig 2017] 及其后续研究中，研究者探索了利用数据中的历史会话作为依据，预测与当前会话最相似的下一个事件。尽管这种方法在概念上简单，却证明了其性能能够与复杂得多的模型相媲美，甚至更胜一筹。最后， [Quadrana 等人 2018], 还提出了多种替代方案，包括依赖（非神经网络）图模型的方法，以及采用离散优化技术的方法；不过，考虑到可推荐物品通常数量庞大，后一类问题可能会受到可扩展性问题的影响。

2.1.3 序列式与会话式推荐系统的评估

在实际应用中，推荐系统会根据其应实现的组织目标和宗旨（例如，提高客户留存率），以及为实现这些目标而执行的具体计算任务（例如，帮助用户发现新内容）进行评估 [Jannach and Adomavicius 2016]。序列感知推荐系统的总体目标通常与传统推荐系统并无太大差异，但序列式与基于会话的推荐器往往会根据其输入数据的特定性质来实现若干计算任务，正如 [Quadrana et al. 2018] 所指出的：

- 情境适应是基于会话的算法所面临的基本问题，即根据当前会话中观察到的交互，推测用户的情境和意图。因此，观察到的用户行为构成了一种交互式上下文 [Adomavicius et al. 2022]。

- 趋势检测 仍是学术研究中一个尚未得到充分探索的课题，既涉及个体趋势（例如兴趣漂移），也涉及学界趋势；后者可能取决于季节性模式或物品的流行度峰值。

•重复推荐通常不会在矩阵补全问题中予以考虑，但对于消耗品（例如打印机墨水）而言，它们可能很有价值；此外，也可以借此提醒用户他们过去喜欢过的物品（例如曾经喜爱的艺术家）。

•顺序约束与模式的考量：在许多领域中，物品的推荐顺序有严格要求，必须按照特定顺序推荐（例如学习课程推荐）；在其他领域中，事物可能有一种“自然”的推荐顺序（例如电影续集）。序列感知推荐系统可以将学习和/或强制执行此类模式作为一项计算任务。

原则上，第 1.4 节概述的所有评估方法都可用于评估序列式与会话式推荐系统，包括离线实验、用户研究和现场测试（A/B 测试）。离线实验在学术研究中占据主导地位，序列感知推荐系统领域也不例外；与往常一样，这类实验旨在同时抽象价值视角和具体的计算任务。

2.1.3.1 离线评估

请回想前文：序列感知推荐系统通常会输出物品排序列表。按照离线实验中常见的“隐藏并预测”评估方法，可以采用信息检索中用于衡量排序准确率的常见评估指标，例如精确率和召回率。因此，序列感知推荐系统的评估流程包括：隐藏给定事件序列中的若干交互，随后衡量算法预测隐藏交互的效果。

在传统矩阵补全问题设置的评估中，隐藏元素（即测试集）的选择通常采用随机方式，例如随机选取 20% 的数据用于测试。此外，许多情况下还会采用交叉验证，以提高对观测结果的置信度。这种完全随机的方法对序列感知推荐问题并无意义，因此必须采用其他流程。

数据划分。 将数据划分为训练集、验证集和测试数据集时，必须考虑时间顺序，并且要预测的留出交互必须发生在用于训练的交互之后。数据划分过程有多种策略，以下是较为常见的几种。

• 在序列问题设置中，通常会按用户划分数据。常见做法是隐藏每位用户最后一次交互，并将其放入测试集。[1]

[1]虽然这是文献中常用的方法，但它可能违反这样一项原则：用于训练的其他用户交互事件，实际上发生在针对特定用户所预测的交互之后。

- 相比之下，在基于会话的场景中，更常采用基于时间的划分方式。研究数据集通常涵盖较长的时间段，例如几周。测试时，常使用最后一天或最后几天的会话（并将此前一天或几天用于验证）。

进行测量。在序列推荐问题中，如果隐藏每位用户的最后一次交互，那么进行测量就是要确定算法是否将隐藏交互排入前 n 列表，以及将其排在什么位置。因此，常见的评估指标包括给定截断长度下的命中率、归一化折损累计增益或平均倒数排名。

对于基于会话的问题，也可以应用这些指标，但评估流程略有不同。请记住，测试数据集通常包含完整的用户会话，而基于会话的推荐要解决的具体问题，是根据会话开端预测下一次交互。文献中报告的方法包括“揭示”会话中除最后一次交互以外的所有交互、揭示前k 个元素，或在测量时逐步揭示交互。应用准确率指标的方式也有不同变体。例如，可以使用命中率、归一化折损累计增益和平均倒数排名等典型指标，在给定截断阈值下衡量算法预测会话中紧接着的下一件物品的表现。或者，也可能更为重要的是，衡量推荐器返回的前 n 列表中包含了会话后续（且被隐藏的）多少次交互。在这种情况下，通常会使用精确率和召回率等指标。

请注意，与传统推荐场景一样，除了准确率之外，也可以考虑其他质量因素。文献中特别关注的准确率之外的方面包括推荐列表的多样性、目录覆盖率、流行度偏差或可扩展性。

一般来说，请注意，虽然我们在数据中隐藏了交互，但通常要对物品进行预测。如果日志中只有一种类型的交互，例如音乐流媒体网站上的收听事件，那么物品预测就相当于预测收听事件。然而，现实世界的数据集通常包含不同类型的交互，例如电子商务网站上的物品浏览、加入购物车操作或购买行为。在许多学术研究中，只考虑最常见的一种交互类型，例如物品浏览。但在现实中，预测物品购买行为可能重要得多，而且应使用所有可用的交互类型进行训练与预测。

交叉验证。 鉴于数据具有序列特性，基于随机划分进行交叉验证并无意义。在文献中，通常只采用一次训练测试划分，但这存在一定风险：所得结果可能仅适用于该次划分。因此，另一种方法是对数据进行多个相互重叠或不重叠的时间划分，并在这些划分上重复测量，参见 [Ludewig 等人2021]。

2.1.3.2 以用户为中心的评估

尽管以用户为中心的研究在一般的推荐系统文献中并不少见，但在序列推荐和基于会话的推荐领域仍十分罕见。在一项音乐领域的研究中 [Kamehkhosh 和 Jannach 2017],研究人员比较了不同的下一曲音乐推荐技术。在该研究中，参与者看到给定播放列表的若干种备选续播方案。在另一项研究中 [Ludewig 和 Jannach 2019], 研究人员创建了一个在线互动广播电台，用于实验。系统采用不同的算法，根据用户提供的起始曲目以及广播收听会话期间的反馈来创建播放列表。这两项研究都对用户的质量感知提供了有趣的见解，并探讨了离线准确率指标在多大程度上与用户感知一致。除了这类特定的研究设置外，还可以采用通用框架对推荐系统进行以用户为中心的评估，例如 [Pu 等人 2011] 或 [Knijnenburg 等人 2012]。

2.1.3.3 真实世界评估

对于较为传统的推荐问题，推荐系统在真实世界部署中已有不少成功案例，参见 [Jannach and Jugovac 2019]。而在序列推荐与基于会话的推荐领域，这类报告则十分少见。一个例外是 [Kouki 等人 2020],的研究，该研究报告了在家居装修领域的一家在线商店中使用基于会话的推荐技术的情况。最终，与现有的零售推荐商业软件相比，现代神经网络方法显著提升了相关企业关键绩效指标。[2]

这项工作另一个有趣之处在于，研究者如何从若干现有的基于会话的算法中决定采用哪一种。首先，作者通过离线评估对多种技术进行了基准测试，其中既包括概念简单的方法，也包括较为复杂的方法。 [Ludewig 等人2021], 在复现先前的离线实验时发现，按照典型的准确率指标衡量，简单方法在许多情况下优于最新的神经网络技术。然而，随后由人类专家进行的评估表明，在给定的应用领域中，表现良好的最近邻技术未必是最佳选择。虽然这种技术通常能够正确预测一个被隐藏的元素，但其他算法在推荐多个可能相关的物品方面表现得更为出色。尤其是，其他算法往往会返回与被隐藏物品相似的物品，因此这些物品可能成为客户的购买替代品。因此，这项研究不仅强调了离线评估可能存在的局限性，也指出在实践中，了解特定应用的具体情况，以及推荐器应如何创造价值，十分重要。

[2]遗憾的是，我们并不了解这一先前系统的内部工作机制；例如，不清楚它实现的是基于会话的方法还是传统方法。

2.1.4 讨论与展望

尤其是，基于会话和会话感知是极具实际意义的问题，近年来我们已经看到针对这些问题提出了许多技术方案。[3] 然而，尽管人们开发了许多复杂的算法，令人惊讶的是，在离线评估中，许多情况下基于最近邻的非常简单的技术也具有竞争力，甚至优于最新的神经网络技术 [Garg 等人，2019，Kouki 等人，2020，Ludewig 等人 2021]。不过，这种现象并非基于会话的推荐问题所独有；请参见[Ferrari Dacrema 等人 2021]对传统 top-N 推荐场景中观察到的类似问题的讨论。总的来说，这表明该领域可能存在某些阻碍进展的方法论问题。

另一方面，这些观察也表明，基于会话的推荐和序列推荐算法仍有巨大的改进潜力。尤其是，利用侧信息（例如物品元数据）、利用上下文信息，以及同时考虑多种类型的交互，似乎都是值得在未来研究的方向。

2.2 推荐中的流行度、偏差与重复出现

随着推荐的视角从评分预测转向物品排序，人们观察到一种新现象：离线评估往往会奖励对热门物品的推荐，也就是说，这些物品经常与用户发生交互 [Cremonesi et al. 2010]。当物品的数据分布——即所谓的流行度分布——呈现明显的长尾形态时，这种现象就会造成强烈的偏差。在随机数据划分中，流行度对评估的影响显然是一种自我实现的预言：训练数据最多的物品——热门物品——也拥有最多的测试数据，而这些数据会被用作相关性判断，正如第 1.4.3 节所述。因此，在推荐中将热门物品排在较高位置，可以减少排序列表顶部的缺失评分，从而提高获得较高相关性指标值的机会 [Bellog´ín et al. 2017, Ca˜namares and Castells 2018]。即使采用时间划分，划分点两侧的流行度分布通常也存在相关性（因为物品流行度会随时间延续），因此流行度偏差依然存在 [Bellog´ín et al. 2017, Mena-Maldonado et al. 2021]。

鉴于这种效应，最先进的协同过滤算法也严重偏向于推荐热门物品，这似乎并不令人意外，因为研究人员已系统地发现了这一点 [Ca˜namares 和 Castells 2017年，Jannach 等人，2015b]。这一认识推动了两个方向的研究：a) 避免或抵消评估和算法中的偏差，这隐含地将偏差视为可能扭曲学习与评估的因素[Bellog´ín etal. 2017年，Jadidinejad 等人，2021年，Steck 2010, 2011]；b) 更好地理解偏差可能源自何处，并探讨它是否可能无害，甚至有用

[3]有关近期更详细的讨论，另请参阅 [Jannach et al. 2022]。

在一定程度上 [Ca˜namares 和 Castells 2018]。另一个根本性视角近期也日益受到关注：它认为，推荐具有双重作用，一方面满足用户需求，另一方面从用户与所提供推荐的交互中学习。我们将在下文简要讨论这些视角。

2.2.1 偏差消除

从某种角度来看，如果从附加价值的视角出发，流行度本身可能并不受欢迎，因此应当避免。这正是将新颖性视为推荐的补充目标的观点，如第 1.4.5.1 节所述。即使从这一视角出发，新颖性仍应与相关性结合，而相关性也需要评估；追求相关性的部分及其评估可能会受到流行度偏差的影响。已有研究提出了具体的评估程序，旨在消除指标中的流行度偏差 [Bellog´ín 等人，2017年，Schnabel 等人，2016年，Steck 2010, 2011] ，以及算法中的流行度偏差 [Hern´andez-Lobato 等人，2014年，Jannach 等人，2017，Liu 等人， 2020]。其中，针对离策略学习与评估的反事实推理方法正在得到积极研究，这些方法采用逆倾向评分（IPS）等技术 [Jadidinejad 等人，2021年，Schnabel 等人，2016年，Swaminathan 等人， 2017]。IPS 通过对用户-物品空间中的观测分布进行建模，并在评估指标和学习算法中对其进行补偿（即根据观测到偏好的概率降低其权重），为指标和目标函数生成无偏估计量。在此背景下应用 IPS 面临的挑战包括：如何根据有偏样本对倾向偏差进行建模，以及当观测概率可以无限趋近于零时，逆倾向评分权重的方差可能变得无界。

2.2.2 理解偏差

除了具备消除或利用偏差的能力之外，一些作者还主张，应更深入地理解流行度分布，以及它们作为评估中扭曲因素可能造成的危害或无害影响 [Ca˜namares and Castells 2018]。毕竟，推荐系统应用会有意利用流行度信号 [Amatriain and Basilico 2015], , 因为流行度未必完全是坏属性；而在评估中天真地抵消流行度，则可能会把这一信号视同随机推荐。Ca˜namares and Castells[2018]对流行度分布的形成进行了理论探讨，并找出了决定在存在流行度偏差时，系统比较结果能否在评估中保持不变的特定条件。答案的关键在于观测、相关性和物品作为随机变量之间的条件依赖关系，这解释了为什么多数选择和多数偏好有助于提供有效推荐。随后，研究者通过实验扩展了这些发现，表明在有偏差和无偏差的数据上进行评估时，结果具有很高的一致性 [Mena-Maldonado et al. 2021]。换言之，这一研究方向的总体发现是，反馈循环和马太效应 [Merton 1968]有时能够形成良性

组成部分 [Ciampaglia etal. 2018]。基于这一认识，Zhang et al. [2021] 探讨了在训练和推理阶段更精细地处理流行度的方法，旨在利用其有用信号的同时减轻扭曲。

2.2.3 反馈回路

离线数据中的一个主要偏差来源，是收集交互数据（即推荐输入数据）的生产系统。推荐系统通常集成在多组件应用中，用户可以通过不同途径与物品交互并提供反馈：搜索、浏览、推荐、首页等。每个此类组件都会引入自身的偏差。如果推荐组件会利用用户与其自身输出之间的交互数据，就会形成反馈回路 [Chaney 等人 2018b, Fleder and Hosanagar 2009]。这种回路与离线评估相结合，会产生放大效应（也称雪球效应、马太效应或强者愈强效应）：为评估而收集的数据会奖励并强化生产系统中根深蒂固的假设，使热门物品愈发热门，并抑制改变 [Chaney 等人 2018a]。因此，系统可能会困在自身的假设中，错失那些隐藏在未观察到的选择中的价值与机会。

此外，这类循环可能产生不良激励：推荐策略盲目地以用户参与为目标，而用户参与有时与短期收入紧密相关——推荐可能因此诱导强迫性行为或有害行为，而不是（或不只是）服务于建设性目的 [Whittaker et al. 2021]。这种兔子洞陷阱充其量可能导致长期业务表现不佳，最坏情况下则可能无意中助长错误信息传播、极化和激进化等有害现象。预防此类风险和有害影响并不简单，而且在很大程度上需要持续监控和直接的专业干预、修订商业模式与激励机制，并应对复杂的伦理问题 [Stray 2020, Zuckerberg 2018]。

从更一般化和算法化的角度来看，打破反馈循环的一种有前景的方法是强化学习视角：在这一视角下，算法被理解为一个同时旨在取悦用户并进行学习的智能体 [Sutton and Barto 2021]。推荐被视为将用户兴趣与当前系统知识中最匹配内容进行匹配的机会（即时优化），同时也通过探索尚未知晓或不确定的领域来扩展这些知识，从而在未来产生更好的推荐。最优解同时关注长期和短期目标（探索与利用），力求实现微妙的平衡。毕竟，推荐系统的使命并不局限于为每位用户提供单次推荐，而是希望在较长时间跨度内，与客户建立一种反复、循环且互惠互利的关系。然而，该领域的大量文献却隐含地建立在与用户仅有一次性互动的观点之上。

该领域的发展采用了所谓的多臂老虎机（MAB）方法[Sutton 和 Barto 2021], ，其中物品被视为待选择的动作（即待推荐的选项）

[Hill 等人 2017]。动作被选中后，会返回一个奖励（即用户的效用或价值），其取值来自未知分布；系统的目标是在一段时间内最大化累积奖励。多臂老虎机算法通过引入可控程度的引导式随机探索，打破反馈回路，帮助跳出局部极大值（简单来说，就是加入少量受控且有依据的随机性），从而比传统的纯利用方法（包括第 1.3 节讨论的所有算法）更好地平衡探索与利用。为定义个性化解决方案，人们开发了所谓的多臂老虎机上下文变体，其中上下文从根本上包含用户，奖励则取决于上下文 [Li 等人 2016]。

对上下文多臂老虎机的进一步推广，会考虑推荐对用户状态产生的影响，从而将用户与系统的交互表示为马尔可夫决策过程（MDP） [Xin 等人 2020]。MDP 也很适合用于表示序列推荐 [Xin 等人，2022a], ，我们已在第 2.1 节讨论过这一点。基于 MDP 的解决方案通常需要大量在线交互才能发挥效用，而要开始提供优质推荐，可能需要耗费难以承受的时间 [Afsar 等人 2022]。可以通过利用更丰富的记录交互数据进行补充离线训练来缓解这一问题，但这会引入日志策略带来的偏差 [Xin 等人，2022b],，需要按照前文所述加以处理。推荐中的强化学习是一个快速发展的领域，与反事实学习和评估密切相关；读者可从 Afsar 等人的综述入手，进一步深入阅读 [2022]。

最后，从更广泛的抽象视角来看，流行度偏差可以被视为源于一个复合反馈回路：其中涉及用户—物品发现与反馈的多种渠道，这些渠道相互作用并彼此强化。直接的新颖性增强、偏差对抗技术和强化学习，可以被视为朝着共同方向寻求推荐中一定程度的健康探索的多种解决方案。

2.3 推荐系统的影响与价值

过去三十年来，推荐系统已成功应用于许多领域，其对消费者和企业的价值毋庸置疑 [Jannach and Jugovac 2019]。然而，与应用机器学习的许多其他领域（包括信息检索的各个子领域）一样，推荐系统的学术研究大多并非基于已部署系统开展，而是主要依赖以数据为中心的方法。此外，与基于观测和分析的数据研究不同，后者的目标例如是从记录数据中理解某些现象，而推荐系统研究几乎完全专注于开发新型预测模型。因此，这类实验中的研究问题通常不是“是什么让一条推荐取得了成功？” [Jannach et al. 2017] ，也不是“热门物品被推荐后，用户是否更

有可能购买它们？”，而几乎总是关注新的预测模型能否更准确地预测留出的过去数据。

因此，如今推荐系统研究的终极目标是提高算法的预测准确率。其依据是一种普遍看来合理的假设：准确率越高，推荐列表中排名靠前的位置就越可能出现更多相关项。这样一来，用户便能更容易发现这些相关项，从而减少信息过载。在这一背景下，几乎所有推荐系统的一个核心特征都在于对物品推荐进行个性化 。[4] 因此，推荐系统可以被视为实施数字化一对一营销和超细分的策略。

遗憾的是，过去数据上的预测准确率更高，就能为用户或推荐提供方带来更高整体价值 这一假设，两者之间的对应关系并不确定。例如，许多学术研究报告指出，较高的预测准确率有时才会提升用户对质量的感知 ，这一结论来自用户研究。产业界的多项研究也报告称，例如 Netflix 的研究表明，离线实验往往无法预测线上成功 [Gomez-Uribe and Hunt 2015]。仅仅依赖准确率指标的局限性不难说明。设想一位用户喜欢前两部《复仇者联盟》电影，并在某个在线视频租赁平台上给它们打了高分。随后，某个算法向该用户推荐另外三部《复仇者联盟》电影以及相关的超级英雄电影，甚至可能达到 100% 的准确率。尽管准确率达到 100%，这项推荐对用户和平台提供方而言可能都价值不大：对用户来说，推荐列表显而易见，而且可能过于单调；对平台提供方来说，用户反正也会租这些电影。因此，推荐这些电影甚至可能错失了推广其他物品的机会。

鉴于这些观测，推荐系统领域亟须超越仅以离线实验和抽象准确率指标为基础的研究。这一转变尤为重要，特别是因为，对于给定数据集而言，某项准确率指标的小幅提升——这类结果在已发表的研究中十分常见——在实践中是否真的有意义，尚不明确。

因此，在本节中，我们首先强调在为推荐系统提出新算法或设计用户界面时，考虑人在回路中的重要性（第2.3.1节）。随后，我们将探讨推荐系统为消费者和推荐服务提供商创造价值的各种方式，以及根据预期目的评估推荐系统为何十分重要（第 2.3.2 节）。

[4]虽然在信息检索场景中，搜索结果的个性化也已得到广泛研究，但通常认为，非个性化推荐（例如推荐普遍热门的物品）无法与个性化推荐相提并论。

2.3.1 了解推荐对人在环中的影响 p

在实践中构建推荐系统时，可能需要考虑一些相关问题，例如：

- 用户会认为系统提供的推荐有用吗?
- 他们会认为这些推荐新颖、多样或有趣吗？最终，他们会满意 这些推荐吗?
- 他们会相信系统的推荐值得信赖吗？他们会希望看到解释吗?
- 未来他们会在决策中考虑系统的推荐吗?

所有这些方面都可能影响推荐系统最终的成功，而这取决于用户是否采纳系统的推荐。遗憾的是，典型的离线实验无法回答这些问题。归根结底，在线推荐在很大程度上是一个人机交互问题，即设计一个计算机化的交互式系统，以支持用户完成信息查找或决策任务。因此，涉及用户的研究是探索上述问题的必要手段。

这类研究最常见的形式是实验，通常采用随机对照试验的形式：研究参与者会被随机分配到一个或多个试验组，或控制组。在许多此类研究中，参与者会在受控环境（通常是实验室）中与原型系统的不同版本进行交互。例如，在一项研究向用户提供解释所产生影响的实验中，一组参与者会与提供解释的系统进行交互，另一组则与不提供解释的系统进行交互。因此，是否提供解释就是实验中被操纵的自变量。根据具体的研究问题，可以测量各种因变量。这些因变量既可以通过客观方式测量，例如记录参与者的行为，如鼠标点击或交互时间；也可以通过问卷请参与者自我报告其感受来获取。例如，如果研究问题是解释是否有助于减少用户的决策努力，既可以测量用户从给定的推荐中作出选择所需的时间，也可以询问参与者对任务感知到的努力程度。实验结束后，可以进行各种统计分析，以评估是否提供解释可能导致了关于决策努力的观测结果，或至少与这些结果相关。

与典型的离线实验不同，这类用户研究以明确的研究问题或假设为指导，而这些问题或假设通常源自理论考量。因此，每项用户研究都需要采用特定的实验设计，以适合回答这些研究问题。此外，研究还需要招募能够代表至少一部分目标人群的参与者。正因如此，用户研究

通常比离线实验更费力，因为离线实验往往并非基于理论考量或明确的研究问题。

正如 [Pu 等人 2011] 或 [Knijnenburg 等人2012] 所提出的，以用户为中心的评估框架是将此类用户实验至少在一定程度上标准化的一种方式。这些框架包含若干通用质量因素，以及推荐系统对用户行为可能产生的影响。图 2.3 概述了 ResQue 框架[Pu 等人 2011], 。该框架部分受技术接受模型的启发，提出评估用户质量感知对其信念、态度和未来行为意图的潜在影响。除了确定模型中的相关构念（变量）外，该框架还提供了一组问卷条目，可用于在实验中测量参与者的主观感知。

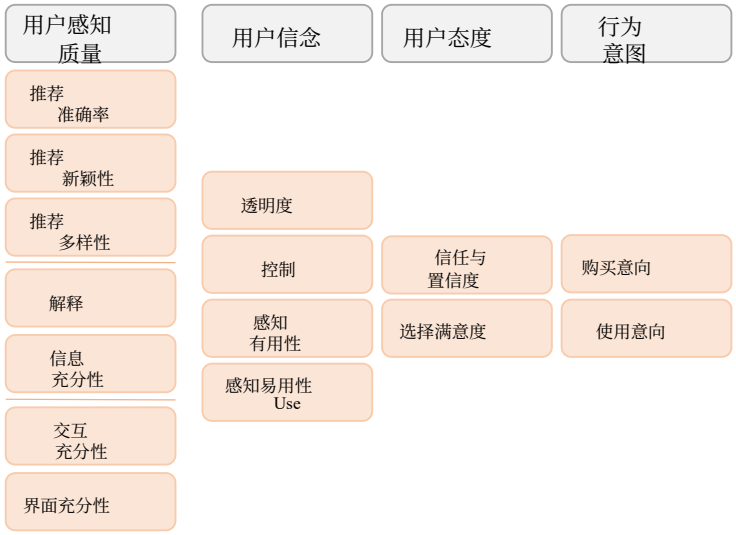


图 2.3 用于推荐系统以用户为中心的评估的 ResQue 框架概览（见 [Pu 等人 2011]）。

总体而言，在信息系统文献中，针对用户开展的受控实验比推荐系统领域计算机科学文献中的此类实验常见得多。鉴于许多重要问题——尤其是与推荐器的用户界面相关的问题 [Jugovac 和 Jannach 2017]——无法通过以改进预测模型为主、但范围狭窄的研究重点来回答，因此推荐系统研究更应考虑开展有人参与的评估。显然，用户研究也有其局限性，例如，研究参与者并未与真实系统交互，或通常不会做出真实的购买决策。尽管如此，这类研究仍能为我们提供相关线索，帮助了解现实世界中哪些因素可能对用户很重要。

最后请注意，受控实验并非有人参与的评估的唯一形式。学术界和产业界常采用其他各种观测式研究，以及定性研究方法，例如焦点小组或访谈。

2.3.2 推荐系统的消费者价值与企业价值

归根结底，任何推荐系统的设计目的都是为一个或多个相关利益相关者创造效用或价值。至少在隐含层面，学术文献历来侧重于消费者价值。一个常见的潜在假设是，如果推荐系统能够为消费者创造价值，例如帮助他们发现新颖物品，那么这至少会间接为提供推荐的企业或组织带来价值。比如，对于基于固定费率订阅的音乐或视频流媒体服务而言，这类考虑似乎是合理的。消费者对服务的参与度越高——得益于不断发现新内容——他们在月底续订订阅的可能性就越大。然而，提供方的目标并不总是与消费者的价值视角完全一致，从而导致多利益相关者推荐问题。

2.3.2.1 将推荐作为多利益相关者优化问题

不同类型的实体都可能参与推荐过程 [Abdollahpouri etal. 2020a, Jannach andBauer 2020]。

- 消费者或最终用户是推荐的接收者。他们可以是个人用户，也可以是用户群体。在后一种群组推荐场景中，参见 [Masthoff and Deli´c 2022], ，通常必须权衡不同群组成员的需求和偏好，也就是说，群组中可能存在多个利益相关者。

- 推荐服务提供商是掌控推荐系统并将其作为业务一部分运行的组织。典型示例包括大型在线零售商或媒体流媒体服务提供商。

- 供应商是创建或提供通过推荐服务销售的物品的实体。在电子商务示例中，他们可能是制造商；在媒体领域，他们可能是艺术家、唱片公司或其他内容提供商。请注意，推荐服务提供商也可能是其所提供物品的供应商。

- 社会：如果某项推荐服务影响到社会中的很大一部分人，例如社交媒体网站上的信息流推荐，那么推荐系统也可能产生社会影响，例如，推荐系统以不平衡的方式强化某些政治观点。

请看以下示例，它说明了考虑多个利益相关者利益时可能面临的挑战 [Jannach 和 Bauer 2020]。设想有一家在线酒店搜索服务，会向酒店业主收取通过该网站完成的每笔 Booking 的佣金。从消费者的角度来看，最佳推荐应满足他们的需求，通常还要符合一定的预算约束。另一方面，推荐服务提供商可能会试图将短期利润（佣金）最大化。因此，推荐系统可能会选择与消费者需求合理匹配、但能为平台带来更高利润的物品。同时，酒店平台也可能试图确保所有酒店供应商都能时不时获得推荐，以维持平台对供应商的吸引力。最后，连锁酒店可能希望推荐系统通过推荐优先推送某一部分产品，例如价格较高的产品。

如今，多利益相关方推荐系统领域的研究仍然有限。例如，已有研究聚焦于互惠推荐 [Palomares 等人 2021],、价格与利润感知推荐系统 [Jannach 和 Adomavicius 2017], ，或算法公平性与偏差领域，例如 [Abdollahpouri 等，2020b、Mehrotra 等，2018b]。总体而言，多利益相关方推荐面临的一项特殊挑战在于，最常见的评估方法——基于历史数据集的准确率评估——通常不足以研究这些问题。因此，需要采用其他方法，例如基于仿真技术的方法 [Ghanem 等，2022、Zhang 等人 2020]。这一点尤为重要，因为平衡多个利益相关方的目标往往需要进行纵向分析，以揭示其对相关利益相关方的长期影响。

2.3.2.2 以影响和价值为导向的推荐系统评估

在实际部署中，推荐系统的表现通常根据由相关组织部门选定或定义的关键绩效指标（KPIs）来确定。实际测量的内容主要取决于希望通过推荐系统实现的目标。因此，关键绩效指标的具体选择或设计也取决于提供方的具体商业模式。

对推荐系统实际成功案例的综述 [Jannach and Jugovac2019] 指出，实践中采用的测量指标主要有以下几类：

- 点击率（*CTR*）：点击率传统上最适用于广告型商业模式，即页面展示量至关重要的场景。然而，在推荐场景中，单纯优化点击率在许多应用中可能过于短视。

- 采用与转化指标：这些指标不止统计点击事件。例如，只有在有迹象表明推荐确实对消费者有用时，才将其视为成功，比如消费者观看了推荐的电影。

- 销售额和收入：这是最直接的衡量方式，例如，可以确定推荐器带动了多少额外销售额。
- 对销售分布的影响：此类指标有助于评估推荐对消费者行为的影响，例如，与没有推荐器的情况相比，推荐是否能够说服消费者选择不同的选项。

- 用户参与度：这项指标通常被认为与长期客户忠诚度有关。当消费者更多地与某项服务交互时，这被视为一种迹象，表明他们会在未来再次购买，或继续订阅或续订订阅。

如前所述，关键绩效指标的具体选择可能与推荐服务提供商所处的具体市场情况及其各自的商业模式密切相关。学术研究通常不会针对这些具体情况的特殊性，而是旨在提供可推广的见解。尽管如此，推荐系统能够为消费者和提供者创造何种价值，也是学术界关注的重要问题。

如上所述，当今的研究在很大程度上（隐含地）聚焦于消费者价值，这通常被概括为帮助用户找到感兴趣的物品（或“找到优质物品” [Herlocker 等人 2004]）。通常，研究会使用预测准确率指标作为衡量这种价值的替代指标，具体做法是评估算法预测用户与物品之间此前记录并留出的正向交互的能力（见第 1.4 节）。尽管这种抽象通常很有用，也为研究问题的操作化提供了标准化方式，但正如上文所述，它可能过度简化问题，甚至误导我们的研究工作。实际上，推荐系统可以通过许多不同方式创造价值，见 [Jannach 和 Bauer 2020]。

为了克服这种略显狭隘的研究操作化，研究人员提出了一个用于推荐系统评估的目的与价值导向框架[Jannach and Adomavicius 2016]。这一概念框架的主要目标，是指导学术界和产业界的研究人员如何衡量推荐系统的成功与否。具体而言，该框架旨在确保衡量内容（即所采用的关键绩效指标或指标）与系统的预期目的相一致。

所提出的框架兼顾消费者视角和提供者视角，包含四个层次：*(i)* 总体目标、*(ii)* 推荐系统的目的、*(iii)* 计算任务、*(iv)*评估方法。为了理解该框架的运作方式，我们以在线媒体流媒体服务为例。从商业视角来看，总体目标 可能是确保该服务的长期盈利能力，而这完全可以独立于任何推荐技术。为实现这一目标，该上下文中推荐器的具体目的 可能是提高用户参与度。实现用户参与度提升的一种方式，是向消费者提供由熟悉内容和新颖推荐组成的搭配，以促进发现。推荐系统的计算任务

因此，是确定这样的搭配。最后，必须选择与最初的总体目标相一致的评估方法 。在这个例子中，可以采用混合方法：*(a)* 通过离线评估衡量不同算法的准确率和新颖性水平；*(b)* 开展用户研究，评估潜在消费者未来使用该服务的意愿；*(c)* 开展实地研究，确定推荐对用户媒体流活动的影响。

在此上下文中最重要的一点——使用所提出的框架时，这一点应该会显而易见——是，仅依赖计算指标（例如预测准确率）可能会受到过多限制。在实践中，准确率等指标与客户留存或客户满意度等更具实际意义的方面严格相关的情况，恐怕并不多。因此，正如 [Jannach 和 Bauer 2020]，更深入讨论的那样，我们似乎需要转变评估推荐系统的范式。这种范式转变最终应能推动该领域产生更具影响力的研究。为实现这一进步，研究人员可以采取以下做法：

- 更常采用所述的目标导向评估方法，确保目标与评估指标更好地保持一致；
- 考虑采用多方法研究方法，其中也包括人在回路评估，既通过受控实验开展，也通过定性研究开展；
- 扩充方法学储备，同时改进基于数据的评估，例如考虑采用基于仿真的方法来考察纵向效应；
- 在适当情况下更常考虑领域与应用特性，并认识到不存在“最佳模型”，同一组推荐在某种情境下可能效果良好，在另一种情境下却未必理想。

遵循这样的建议，最终或许有助于避免算法研究陷入死胡同。算法研究不仅过度关注抽象指标，也正面临某些方法学问题 [Ferrari Dacrema et al. 2021]，这些问题可能会限制当前推荐器研究在实践中的影响。

2.4 总结与未来方向

本文最后将简要总结全文，提供延伸阅读建议，并展望该领域重要的未来方向。

2.4.1 总结

如今，个性化推荐已成为许多在线服务的核心组成部分，文献中也报道了许多成功案例，说明此类系统如何为用户和组织创造价值。由于在实践中取得了广泛成功，过去二十五年来，相关研究兴趣也持续增长。最终，这

促成了一个独立的推荐系统研究领域的形成。该领域在历史上与信息检索、机器学习和人机交互有着深厚的渊源。本文首先从抽象且独立于应用的层面阐述了推荐问题，随后回顾了将这一研究问题具体化为矩阵补全问题的最常见方法。在讨论最常见的推荐算法类型后，我们进一步阐述了如何在“实验室”和“真实环境”中评估推荐系统。最后，我们回顾了基于会话的推荐领域近期的技术进展，以及基于强化学习的方法。

2.4.2 延伸阅读与未来方向

本文主要关注推荐系统的算法层面，而传统上这也是最活跃的研究领域。然而，在实践中构建推荐系统，所需的远不止一个巧妙的算法 [Jannach 等人 2016年b, Xiao 和 Benbasat 2007]。与当下大量的算法研究相比，推荐系统的用户体验问题——即人机交互视角——以及推荐器对用户的影响——即信息系统视角——似乎尚未得到充分探索。关于推荐系统特定应用领域的独特性，似乎也需要开展更多研究；这意味着并不存在适用于所有应用的“最佳”算法，而且只有相对于预期目的，才能评估推荐器是否成功。因此，一种基于用户长期偏好来预测电影相关性评分的算法，可能并不适合推荐时效性较短的新闻文章，也不适合在旅游领域根据上下文推荐音乐或兴趣点。有关推荐系统各种应用的更深入讨论，请参阅《推荐系统手册》[Ricci 等人 2022]。

同一本手册还讨论了推荐系统领域的许多其他热门研究主题，这些主题在本文中未予探讨，包括近期算法的发展、情境感知、针对推荐系统的攻击、隐私问题，以及评估过程中的人为因素。在本文剩余部分，我们将讨论少数几个其他颇具前景的主题，并认为它们值得在未来开展更多研究。

对话式推荐系统。 如今大多数在线推荐系统的用户交互模型都相当简单。在应用程序的特定位置，系统会向用户展示推荐内容，而用户通常唯一可采取的操作就是查看或接受这些推荐。[5] 随着近年来自然语言处理的发展、机器学习整体进步，以及语音设备日益普及，更具交互性的信息检索和推荐形式 [Radlinski

[5]只有在某些情况下，系统才会提供其他机制，让用户对推荐内容提供反馈，或对显示的内容施加一定控制 [Jannach et al. 2016a]。

和 Craswell 2017] 近年来重新受到关注。这类对话式推荐系统（CRS）的一种设想是，它们能够像人类一样与用户进行“自然”对话，综述请参见 [Jannach et al. 2021] 。相应地，这类未来系统将能够支持多种用户意图 [Cai and Chen 2019], ，从而支持交互式偏好 elicitation 与修订、解释以及闲聊。过去，这类 CRS 通常基于人工构建的知识（例如有关可能的对话状态的知识）来构建；如今许多方法也将“端到端学习”作为主要方法，即利用人类之间的大型对话集合训练机器学习模型。近年来，这方面取得了显著进展。不过，纯粹基于学习的方法仍有其局限性，预计未来的 CRS 通常会结合显式领域知识、推理知识和学习组件。

推荐系统中的公平性。 随着人工智能和技术呈指数级发展，公平性与伦理问题日益凸显，渗透并改变着现代社会的各个领域——推荐技术也不例外。推荐功能可能决定 Spotify 上一位音乐人的成败、亚马逊上一位卖家的经营状况、通过 LinkedIn 求职的候选人或招聘人员的境遇、Google Scholar 上一位研究作者的影响力、Booking.com 上一位酒店经营者的生意、YouTube、Instagram 或 TikTok 上一位网络红人的人气，也可能影响人们通过在线新闻和在线媒体接触到的政党、世界观或价值体系。作为连接提供者与消费者的中介，推荐系统无论是否情愿，都肩负着难以卸下的责任。

公平性，更广义地说，负责任的人工智能，确实已成为推荐系统研究中亟待解决的问题，这一点也体现在近期研究文献 [Mehrotra 等人，2018a], 、主题演讲 [Baeza-Yates 2020], 、新设立的专门研究出版渠道 [Elish 等人2021], 以及公众讨论中。研究者正在提出衡量公平性的指标，并研究算法及其改进方法，以避免对群体 [Yao 和 Huang 2017],、少数群体或个人造成不公平待遇，无论他们是推荐内容的提供者还是消费者 [Patro 等人 2020]。这一新研究领域兴起后，人们首先认识到，公平性作为一个概念十分复杂，普适性解决方案也难以实现。各项研究一致强调，培养偏见意识并加以理解，是防止或减轻不公平现象的关键一步（也可能是解决方案中相当重要的一部分）。预计未来还将持续涌现大量相关思考与研究。

评估中的离线／在线不一致。 如第 1.4.1 节所述，离线评估结果与在线 A/B 测试的结果往往只有较弱的一致性 [Garcin 等人，2014，Gomez-Uribe 和 Hunt，2015，Jannach and Jugovac 2019]。已知原因可能只能部分解释这一现象，例如指标不匹配、离线数据中的偏差，或在评估技术被整合到复杂的多步骤推荐流水线中时受到外部组件的干扰；而这类流水线通常应用于工业级应用 [Amatriain

and Basilico 2015]。但目前似乎还没有得出更全面的解释，这个问题也可能尚未得到充分理解。正如前文所述，研究者正从不同角度为此付出重要努力 [Gilotte 等，2018，Jannach andBauer 2020]。如果能改进离线评估，使其更接近对推荐最终效果的衡量，将对未来推荐技术的发展产生极大的积极影响。

参考文献

H. Abdollahpouri、G. Adomavicius、R. Burke、I. Guy、D. Jannach、T. Kamishima、J. Krasnodebski 和 L. Pizzato。2020a。多利益相关者推荐：综述与研究方向。用户建模与用户适应交互，30：127–158。H. Abdollahpouri、M. Mansoury、R. Burke 和 B. Mobasher。2020b。推荐中流行度偏差、校准与公平性之间的联系。载于第十四届 *ACM* 推荐系统会议（*RecSys 2020*），第 726–731 页。ACM。P. Adamopoulos 和 A. Tuzhilin。2014年。推荐系统中的意外性：或如何预期意外。推荐系统新颖性与多样性专题，*ACM*《智能系统与技术汇刊》，5(4)。G. Adomavicius 和 A. Tuzhilin。2005年。迈向下一代推荐系统：最先进水平及可能扩展的综述。*IEEE* 知识与数据工程汇刊，17(6)：734–749。G. Adomavicius、K. Bauman、A. Tuzhilin 和 M. Unger。2022年。上下文感知推荐系统：从基础到近期发展。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《推荐系统手册》，3^{rd} 版，第 211–250 页。施普林格，美国纽约州纽约市。D. Afchar 和 R. Hennequin。2020年。利用归因使神经网络可解释：应用于隐式信号预测。载于第 14^{th} *ACM*推荐系统会议论文集，RecSys 2020，第 220–229 页。ACM，美国纽约州纽约市。M. M. Afsar、T. Crump 和 B. Far。2022年6月。基于强化学习的推荐系统：综述。*ACM*计算调查，即将发表。R. Agrawal、S. Gollapudi、A. Halverson 和 S. Ieong。2009年。搜索结果多样化。载于第 2^{nd} *ACM Web* 搜索与数据挖掘会议论文集，WSDM 2009，第 5–14 页。ACM，美国纽约州纽约市。X. Amatriain 和 J. Basilico。2015年。工业界的推荐系统：Netflix 案例研究。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《推荐系统手册》，2^{nd} 版，第 385–419 页。施普林格，美国纽约州纽约市。V. W. Anelli、A. Bellogin、T. D. Noia、D. Jannach 和 C. Pomo。2022年。Top-n 推荐算法：追求最先进水平。载于第三十届 *ACM* 用户建模、适应与个性化会议（*UMAP 2022*）。R. Baeza-Yates。2020年。搜索与推荐系统中的偏差。载于 R. L. T. Santos、L. B. Marinho、E. M. Daly、L. Chen、K. Falk、N. Koenigstein 和 E. S. de Moura 编，14^{th} *ACM*推荐系统会议（*RecSys 2020*），第 2 页。ACM。M. Balabanovi´c 和 Y. Shoham。1997年3月。Fab：基于内容的协同推荐。*ACM*通讯，40(3)：66–72。

D. Bamber。1975年。序数优势图上方的面积与受试者工作特征图下方的面积。数学心理学杂志，12(4)：387–415。R. Bambini、P. Cremonesi 和 R. Turrin。2011年。《推荐系统手册》，章节“面向 IPTV服务提供商的推荐系统：真实的大规模生产环境”，第 299–331 页。施普林格。编者：Francesco Ricci、Lior Rokach、Bracha Shapira 和 Paul B. Kantor。N. J. Belkin 和 W. B. Croft。1992 年。信息过滤与信息检索：同一枚硬币的两面？*ACM* 通讯，35(12)：29–38。A. Bellog´ın、P. Castells 和 I. Cantador。2017年。推荐系统信息检索指标中的统计偏差。信息检索，20(6)：606–634。J. Bennett、C. Elkan、B. Liu、P. Smyth 和 D. Tikk。2007年。KDD杯与研讨会 2007。《*SIGKDD*探索》，9(2)：51–52。D. Billsus 和 M. J. Pazzani。1998年。学习协同信息过滤器。载于第十五届国际机器学习会议论文集，ICML 1998，第 46–54 页。摩根考夫曼出版社。J. S. Breese、D. Heckerman 和 C. Kadie。1998年。协同过滤预测算法的实证分析。载于第 *14th* 人工智能不确定性会议，UAI 1998，第 43–52 页。摩根考夫曼出版社。C. Buckley 和 E. M. Voorhees。2004年。利用不完整信息进行检索评估。载于第 *27th* 年度国际*ACM SIGIR*信息检索研究与发展会议（*SIGIR 2004*），第 25–32 页。ACM，美国纽约州纽约市。R. Burke。2002年。混合推荐系统：综述与实验。用户建模与用户适应交互，12：331–370。W. Cai 和 L. Chen。2019年。迈向会话式推荐中用户反馈意图的分类体系。载于*RecSys' 19*最新成果，第 572–573 页。P. G. Campos、F. D´ıez 和 I. Cantador。2014年。时间感知推荐系统：现有评估协议的全面综述与分析。用户建模与用户适应交互，24(1)：67–119。R. Ca˜namares 和 P. Castells。2017年。基于记忆的协同过滤的概率重构：对流行度偏差的影响。载于第 *40th* 年度国际*ACM SIGIR*信息检索研究与发展会议（*SIGIR 2017*），第 215–224 页。ACM，美国纽约州纽约市。R. Ca˜namares 和 P. Castells。2018年。我应该随大流吗？流行度在推荐系统中效果的概率分析。载于第 *41st* 年度国际*ACM SIGIR*信息检索研究与发展会议（*SIGIR 2018*），第 415–424 页。ACM，美国纽约州纽约市。R. Ca˜namares 和 P. Castells。2020年。关于离线推荐系统评估中的目标项目抽样。载于 *14th ACM*推荐系统会议（*RecSys 2020*），第 259–268 页。ACM，美国纽约州纽约市。R. Ca˜namares、P. Castells 和 A. Moffat。2020年。推荐系统的离线评估选项。《信息检索》，23(4)：387–411。

S. Castagnos、A. Brun 和 A. Boyer。2013年。何时需要多样性……却又不期待它！载于《 *3rd* 信息挖掘与管理进展国际会议论文集，IMMM 2013，第44–50页。葡萄牙里斯本。P. Castells 和 A. Moffat。2022年。离线推荐系统评估：挑战与新方向。《人工智能杂志》，43(2)。P. Castells、N. J. Hurley 和 S. Vargas。2022年。推荐系统中的新颖性与多样性。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《style id='8'>推荐系统手册， *3rd* 版，第603–646页。施普林格，美国纽约州纽约市。O. Celma 和 P. Herrera。2008年。一种评估新颖推荐的新方法。载于《 *2nd ACM*推荐系统会议论文集，RecSys 2008，第179–186页。ACM，美国纽约州纽约市。Y. Cen、J. Zhang、X. Zou、C. Zhou、H. Yang 和 J. Tang。2020年。用于推荐的可控多兴趣框架。载于《 *26th ACM SIGKDD*知识发现与数据挖掘国际会议论文集，KDD 2020，第2942–2951页。ACM，美国纽约州纽约市。A. J. B. Chaney、B. M. Stewart 和 B. E. Engelhardt。2018a。推荐系统中的算法混杂如何增加同质性并降低效用。载于第 *12* 届 *ACM*推荐系统会议论文集，RecSys 2018，第224–232页。ACM，美国纽约州纽约市。A. J. B. Chaney、B. M. Stewart 和 B. E. Engelhardt。2018b。推荐系统中的算法混杂如何增加同质性并降低效用。载于 S. Pera、M. D. Ekstrand、X. Amatriain 和 J. O'Donovan 编，《 *12th ACM*推荐系统会议（*RecSys 2018*）论文集，第224–232页。ACM。L. Chen、G. Chen 和 F. Wang。2015年。基于用户评论的推荐系统：最先进水平。《*User Model User-Adap Inter*》，25：99–154。L. Chen、Y. Yang、N. Wang、K. Yang 和 Q. Yuan。2019年。意外发现如何提升用户对推荐的满意度？一项大规模用户评估。载于《万维网大会论文集》，WWW 2019，第240–250页。ACM，美国纽约州纽约市。G. L. Ciampaglia、A. Nematzadeh、F. Menczer 和 A. Flammini。2018年。算法流行度偏差如何阻碍或促进质量。《*Scientific Reports*》，8。P. Covington、J. Adams 和 E. Sargin。2016年。用于 YouTube 推荐的深度神经网络。载于《 *10th ACM*推荐系统会议论文集，RecSys 2016，第191–198页。ACM，美国纽约州纽约市。P. Cremonesi、Y. Koren 和 R. Turrin。2010年。推荐算法在 Top-N 推荐任务中的表现。载于《 *4th ACM*推荐系统会议（*RecSys 2010*）论文集，第39–46页。ACM，美国纽约州纽约市。S. Dhelim、N. Aung、M. A. Bouras、H. Ning 和 E. Cambria。2022年。人格感知推荐系统综述。《人工智能评论》，55：2409–2454。M. Dong、F. Yuan、L. Yao、X. Wang、X. Xu 和 L. Zhu。2022年。信任感知推荐系统综述：深度学习视角。《*Knowledge-Based Systems*》，249：108954。M. D. Ekstrand、F. M. Harper、M. C. Willemsen 和 J. A. Konstan。2014年。用户对推荐算法差异的感知。载于《 *8th ACM*推荐系统会议论文集，RecSys 2014，第161–168页。ACM，美国纽约州纽约市。

M. C. Elish、W. Isaac 和 R. S. Zemel 编。2021年。*ACM*公平、问责与透明度会议（*FAccT 2021*）。ACM。Y. Feng、B. Hu、F. Lv、Q. Liu、Z. Zhang 和 W. Ou。2020年。*ATBRG*：用于有效推荐的自适应目标行为关系图网络，第2231–2240页。SIGIR 2020。ACM，美国纽约州纽约市。M. Ferrari Dacrema、S. Boglio、P. Cremonesi 和 D. Jannach。2021。对推荐系统研究中可复现性与进展的令人担忧的分析。《*ACM*信息系统汇刊》，39(2)。D. Fleder 和 K. Hosanagar。2009年。大片文化的下一轮兴衰：推荐系统对销售多样性的影响。《管理科学》，55(5)：697–712。P. W. Foltz 和 S. T. Dumais。1992年。个性化信息传递：信息过滤方法分析。《*ACM*通讯》，35(12)：51–60。F. Garcin、B. Faltings、O. Donatsch、A. Alazzawi、C. Bruttin 和 A. Huber。2014年。swissinfo.ch 新闻推荐系统的离线与在线评估。载于《 8^{th} *ACM*推荐系统会议论文集，RecSys 2014，第169–176页。ACM，美国纽约州纽约市。D. Garg、P. Gupta、P. Malhotra、L. Vig 和 G. Shroff。2019年。用于基于会话的推荐的序列与时间感知邻域：Stan。载于《 42^{nd} 国际*ACM SIGIR*信息检索研究与发展会议论文集，SIGIR 2019，第1069–1072页。M. Ge、C. Delgado-Battenfeld 和 D. Jannach。2010年。超越准确率：通过覆盖率和意外发现评估推荐系统。载于《 4^{th} *ACM*推荐系统会议论文集，RecSys 2010，第257–260页。M. Ge、D. Jannach、F. Gedikli 和 M. Hepp。2012年。推荐列表中多样化物品的排列位置所产生的影响。载于 14^{th} 国际企业信息系统会议（*ICEIS 2012*），第201–208页。N. Ghanem、S. Leitner 和 D. Jannach。2022年。平衡推荐系统的消费者价值与商业价值：基于仿真的分析。*arXiv*预印本 *arXiv:2203.05952*。A. Gilotte、C. Calauz`enes、T. Nedelec、A. Abraham 和 S. Doll´e。2018年。推荐系统的离线 A/B 测试。载于《 11^{th} *ACM*国际网络搜索与数据挖掘会议（*WSDM 2018*）论文集，第198–206页。ACM，美国纽约州纽约市。D. Goldberg、D. Nichols、B. M. Oki 和 D. Terry。1992年。利用协同过滤编织信息织锦。《*ACM*通讯》，35(12)：61–70。C. A. Gomez-Uribe 和 N. Hunt。2015年。网飞推荐系统：算法、商业价值与创新。《管理信息系统汇刊》，6(4)。A. Gruson、P. Chandar、C. Charbuillet、J. McInerney、S. Hansen、D. Tardieu 和 B. Carterette。2019年。通过离线评估为播放列表推荐决策。载于《 12^{th} *ACM*国际网络搜索与数据挖掘会议（*WSDM 2019*）论文集，第420–428页。ACM，美国纽约州纽约市。A. Gunawardana、G. Shani 和 S. Yogev。2022年。评估推荐系统。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《style id='47'>推荐系统手册， 3^{rd} 版，第547–601页。施普林格，美国纽约州纽约市。

I. Guy. 2022. 社交推荐系统。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《推荐系统手册》3rd 版，第 835–870 页。施普林格，美国纽约州纽约市。F. M. Harper 和 J. A. Konstan. 2015. MovieLens数据集：历史与背景。《*ACM*交互式智能系统汇刊》，5(4)。J. He 和 W. W. Chu. 2010. 基于社交网络的推荐系统（*SNRS*），第 47–74 页。Springer US。R. He 和 J. McAuley. 2016. 将相似性模型与马尔可夫链结合，用于稀疏序列推荐。载于《16th *IEEE*国际数据挖掘会议论文集，ICDM 2016，第 191–200 页。X. He、L. Liao、H. Zhang、L. Nie、X. Hu 和 T.-S. Chua. 2017. 神经协同过滤。载于《26th 国际万维网会议论文集，WWW 2017，第 173–182 页。J. L. Herlocker、J. A. Konstan、L. G. Terveen 和 J. T. Riedl. 2004. 评估协同过滤推荐系统。*ACM*信息系统汇刊，22(1)：5–53。J. M. Hern´andez-Lobato、N. Houlsby 和 Z. Ghahramani. 2014. 基于非随机缺失数据的概率矩阵分解。载于《31st 国际机器学习会议论文集，ICML 2014，第 1512–1520 页。机器学习研究论文集，英国谢菲尔德。B. Hidasi 和 A. Karatzoglou. 2018. 采用 top-k增益的循环神经网络，用于基于会话的推荐。载于《27th *ACM*国际信息与知识管理会议论文集，CIKM 2018，第 843–852 页。ACM，美国纽约州纽约市。D. N. Hill、H. Nassif、Y. Liu、A. Iyer 和 S. V. N. Vishwanathan. 2017. 一种用于实时多变量优化的高效多臂老虎机算法。载于《23rd *ACM SIGKDD*知识发现与数据挖掘国际会议（*KDD 2017*）论文集，第 1813–1821 页。ACM。T. Hofmann. 2004. 用于协同过滤的潜在语义模型。*ACM*信息系统汇刊，22(1)：89–115。R. Hu 和 P. Pu. 2011. 利用组织界面增强推荐多样性。载于《*16th* 国际智能用户界面会议论文集，IUI 2011，第 347–350 页。Y. Hu、Y. Koren 和 C. Volinsky. 2008. 面向隐式反馈数据集的协同过滤。载于《8th *IEEE*国际数据挖掘会议论文集，ICDM 2008，第 15–19 页。IEEE计算机学会，美国华盛顿特区。N. Hurley 和 M. Zhang. 2011. top-N推荐中的新颖性与多样性——分析与评估。《*ACM*互联网技术汇刊》，10(4)。A. H. Jadidinejad、C. Macdonald 和 I. Ounis. 2021年9月. 推荐系统离线评估中的辛普森悖论。*ACM*信息系统汇刊，40(1)。ISSN 1046-8188。D. Jannach 和 G. Adomavicius. 2016. 目的导向推荐。载于《10th*ACM*推荐系统会议论文集，RecSys 2016，第 7–10 页。D. Jannach 和 G. Adomavicius. 2017. 推荐系统中的价格与利润意识。载于《*ACM RecSys 2017*价值感知与多利益相关者推荐研讨会论文集》。D. Jannach 和 C. Bauer. 2020. 逃离麦克纳马拉谬误：迈向更具影响力的推荐系统研究。《人工智能杂志》，41(4)：79–95。

D. Jannach 和 K. Hegelich. 2009. 移动互联网中推荐效果的案例研究。载于《第*10*届*ACM*推荐系统会议论文集》，RecSys '09，第 205–208 页。D. Jannach 和 M. Jugovac. 2019. 衡量推荐系统的商业价值。*ACM*管理信息系统汇刊（*TMIS*），10(4)：1–23。D. Jannach 和 M. Ludewig. 2017. 当循环神经网络遇上基于会话推荐的邻域。载于《 11*th* 推荐系统会议论文集，RecSys 2017。D. Jannach、M. Zanker、A. Felfernig 和 G. Friedrich. 2010. 推荐系统——导论。剑桥大学出版社。D. Jannach、L. Lerche 和 I. Kamehkhosh. 2015a. 不止于"热门榜单"：利用恰当曲目生成连贯的音乐播放列表续接。载于《 9*th* *ACM*推荐系统会议论文集，RecSys 2015，第 187–194 页。ACM，美国纽约州纽约市。D. Jannach、L. Lerche、I. Kamehkhosh 和 M. Jugovac. 2015b. 推荐器会推荐什么：推荐偏差分析及可能的对策。用户建模与用户适应交互，25(5)：427–491。D. Jannach、S. Naveed 和 M. Jugovac. 2016a. 推荐系统中的用户控制：概述与交互挑战。载于 17*th* 国际电子商务与网络技术会议，EC-Web 2016。D. Jannach、P. Resnick、A. Tuzhilin 和 M. Zanker. 2016b. 推荐系统——超越矩阵补全。*ACM*通讯，59(11)：94–102。D. Jannach、M. Ludewig 和 L. Lerche. 2017. 电子商务中的基于会话的物品推荐：短期意图、提醒、趋势与折扣。用户建模与用户适应交互，27(3–5)：351–392。D. Jannach、L. Lerche 和 M. Zanker. 2018. 基于隐式反馈进行推荐。载于 P. Brusilovsky 和 D. He 编，社会信息访问。施普林格。D. Jannach、A. Manzoor、W. Cai 和 L. Chen. 2021. 对话式推荐系统综述。*ACM*计算调查，即将出版。D. Jannach、P. Cremonesi 和 M. Quadrana. 2022. 基于会话的推荐。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《推荐系统手册》 3*rd* 版，第 301–224 页。施普林格，美国纽约州纽约市。T. Joachims. 2002. 利用点击数据优化搜索引擎。载于《 8*th*ACM SIGKDD知识发现与数据挖掘国际会议论文集，KDD 2002，第 133–142 页。ACM，美国纽约州纽约市。M. Jugovac 和 D. Jannach. 2017. 与推荐系统交互：概述与研究方向。*ACM*智能交互系统汇刊（*ACM TiiS*），7(3)。M. Jugovac、D. Jannach 和 L. Lerche. 2017. 根据个体用户倾向，高效优化多个推荐质量因素。专家系统及其应用，81：321–331。I. Kamehkhosh 和 D. Jannach. 2017. 用户对下一曲音乐推荐的感知。载于《*2017*年用户建模、适应与个性化会议论文集》，UMAP 2017，第 113–121 页。

W. Kang 和 J. J. McAuley。2018。自注意力序列推荐。载于《 18th IEEE国际数据挖掘会议论文集，ICDM 2018，第197–206页。IEEE计算机学会。K. Kapoor、V. Kumar、L. Terveen、J. A. Konstan 和 P. Schrater。2015。“我有时喜欢探索”：适应动态用户新颖性偏好。载于《 9th ACM推荐系统会议论文集，RecSys 2015，第19–26页。E. Kirshenbaum、G. Forman 和 M. Dugan。2012。在 forbes.com 上对个性化文章推荐方法进行实时比较。载于《数据库中的机器学习与知识发现》，第51–66页。B. P. Knijnenburg 和 M. C. Willemsen。2015。通过用户实验评估推荐系统。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《推荐系统手册》，第309–352页。施普林格，美国纽约州纽约市。B. P. Knijnenburg、M. C. Willemsen、Z. Gantner、H. Soncu 和 C. Newell。2012。解释推荐系统的用户体验。《用户建模与用户适应交互》，22：441–504。Y. Koren。2008。因子分解与邻域相遇：一种多面向的协同过滤模型。载于《 14th ACM SIGKDD知识发现与数据挖掘国际会议（KDD 2008）论文集，第426–434页。ACM，美国纽约州纽约市。Y. Koren。2009。具有时间动态性的协同过滤。载于《 15th ACM SIGKDD知识发现与数据挖掘国际会议（KDD 2009）论文集，第447–456页。ACM，美国纽约州纽约市。Y. Koren、R. M. Bell 和 C. Volinsky。2009。推荐系统的矩阵分解技术。《Computer》，42(8)：30–37。P. Kouki、I. Fountalis、N. Vasiloglou、X. Cui、E. Liberty 和 K. Al Jadda。2020。从实验室走向生产：家居装修领域基于会话的推荐案例研究。载于《 14th ACM推荐系统会议论文集，RecSys 2020，第140–149页。W. Krichene 和 S. Rendle。2020。关于物品推荐的采样指标。载于《 26th ACM SIGKDD知识发现与数据挖掘国际会议（KDD 2020）论文集，第1748–1757页。ACM，美国纽约州纽约市。N. Lathia、S. Hailes、L. Capra 和 X. Amatriain。2010。推荐系统中的时间多样性。载于《 33rd 年度国际ACM SIGIR信息检索研究与发展会议论文集，SIGIR 2010，第210–217页。ACM，美国纽约州纽约市。N. K. Lathia。2010。随时间变化的协同过滤评估。博士论文，伦敦大学学院，英国。S. Latifi、N. Mauro 和 D. Jannach，2020。会话感知推荐：对最先进水平的一次出人意料的探索。S. Li、A. Karatzoglou 和 C. Gentile。2016。协同过滤多臂老虎机。载于《39th 国际ACM SIGIR信息检索研究与发展会议论文集，SIGIR 2016，第539–548页。ACM，美国纽约州纽约市。D. Liang、R. G. Krishnan、M. D. Hoffman 和 T. Jebara。2018。用于协同过滤的变分自编码器。载于《万维网会议论文集》，WWW 2018，第689–698页。国际万维网会议指导委员会，日内瓦共和国与州，CHE。

G. Linden、B. Smith 和 J. York。2003。亚马逊推荐：物品到物品协同过滤。《*IEEE Internet Computing*》，7(1)：76–80。B. Liu、J. Bennett、C. Elkan、P. Smyth 和 D. Tikk。2007。KDD杯与研讨会2007。载于《13*th* *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，KDD 2007。ACM。D. Liu、P. Cheng、Z. Dong、X. He、W. Pan 和 Z. Ming。2020。通过均匀数据进行反事实推荐的通用知识蒸馏框架。载于《43*rd* 国际*ACM SIGIR*信息检索研究与发展会议论文集，SIGIR 2020，第831–840页。ACM，美国纽约州纽约市。T.-Y. Liu。2009。信息检索中的学习排序。《信息检索基础与趋势®》，3(3)：225–331。M. Ludewig 和 D. Jannach。2019。面向自动化广播电台的基于会话推荐的以用户为中心的评估。载于《13*th* *ACM*推荐系统会议论文集，RecSys 2019，第516–520页。M. Ludewig、S. Latifi、N. Mauro 和 D. Jannach。2021。基于会话的推荐算法的实证分析。《用户建模与用户适应交互》，31：149–181。B. M. Marlin 和 R. S. Zemel。2009。基于非随机缺失数据的协同预测与排序。载于《3*rd* *ACM*推荐系统会议（*RecSys 2009*）论文集，第5–12页。ACM，美国纽约州纽约市。J. Masthoff 和 A. Deli´c。2022。群体推荐系统：超越偏好聚合。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《推荐系统手册，》3*rd* 版，第381–420页。施普林格，美国纽约州纽约市。S. M. McNee、J. Riedl 和 J. A. Konstan。2006。准确并不足够：准确性指标如何损害推荐系统。载于《*ACM CHI 2006*计算系统中的人为因素会议论文集》，CHI 2006，第1097–1101页。ACM，美国纽约州纽约市。R. Mehrotra、J. McInerney、H. Bouchard、M. Lalmas 和 F. Diaz。2018a。迈向公平市场：对推荐系统中相关性、公平性与满意度之间权衡的反事实评估。载于《27*th* *ACM*信息与知识管理国际会议论文集，CIKM 2018，第2243–2251页。ACM，美国纽约州纽约市。R. Mehrotra、J. McInerney、H. Bouchard、M. Lalmas 和 F. Diaz。2018b。迈向公平市场：对推荐系统中相关性、公平性与满意度之间权衡的反事实评估。载于《27*th* *ACM*信息与知识管理国际会议论文集，CIKM 2018，第2243–2251页。ACM，美国纽约州纽约市。E. Mena-Maldonado、R. Ca˜namares、P. Castells、Y. Ren 和 M. Sanderson。2021。推荐系统评估中假阳性指标的流行度偏差。《*ACM*信息系统汇刊》。R. K. Merton。1968。科学领域的马太效应。《*Science*》，159(3810)：56–63。T. Mikolov、I. Sutskever、K. Chen、G. Corrado 和 J. Dean。2013。词语和短语的分布式表示及其组合性。载于《26*th* 神经信息处理系统国际会议论文集，NIPS 2013，第3111–3119页。Curran Associates Inc.，美国纽约州雷德胡克。

A. Moffat 和 J. Zobel。2008年12月。用于衡量检索有效性的排名偏置精确率。*ACM*信息系统汇刊，27(1): 2:1–2:27。ISSN 1046-8188。C. Musto、M. de Gemmis、P. Lops、F. Narducci 和 G. Semeraro。2022年。语义与基于内容的推荐。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《style id='3'>推荐系统手册，3*rd* 版，第251–298页。施普林格，纽约，纽约州，美国。A. N. Nikolakopoulos、X. Ning、C. Desrosiers 和 G. Karypis。2022年。信任你的邻居：推荐系统基于邻域的方法综合综述。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《style id='8'>推荐系统手册， 3*rd* 版，第39–89页。施普林格，纽约，纽约州，美国。X. Ning 和 G. Karypis。2011年。SLIM：面向Top-N推荐系统的稀疏线性方法。载于《 11*th* *IEEE*国际数据挖掘会议论文集，ICDM 2011，第497–506页。IEEE计算机学会，华盛顿特区，美国。J. Oh、S. Park、H. Yu、M. Song 和 S.-T. Park。2011年。基于个人流行度倾向的新颖性推荐。载于《 11*th* *IEEE*数据挖掘会议论文集，ICDM 2011，第507–516页。S. Okura、Y. Tagami、S. Ono 和 A. Tajima。2017年。面向数百万用户的基于嵌入的新闻推荐。载于《 23*rd* *ACM SIGKDD*知识发现与数据挖掘国际会议论文集，KDD 2017，第1933–1942页。ACM，纽约，纽约州，美国。I. Palomares、C. Porcel、L. Pizzato、I. Guy 和 E. Herrera-Viedma。2021。互惠推荐系统：前沿文献分析、挑战以及迈向社交推荐的机遇。信息融合，69: 103–127。G. P. Patil 和 C. Taillie。1982年。多样性作为一个概念及其测量。美国统计学会杂志，77(379): 548–561。G. K. Patro、A. Biswas、N. Ganguly、K. P. Gummadi 和 A. Chakraborty。2020年。Fairrec：双边平台中个性化推荐的双边公平性。载于万维网会议论文集，WWW 2020，第1194–1204页。ACM / IW3C2。M. J. Pazzani 和 D. Billsus。2007年。基于内容的推荐系统，第325–341页。施普林格出版社，柏林，海德堡。J. Pennington、R. Socher 和 C. Manning。2014年10月。GloVe：用于词表示的全局向量。载于*2014*年自然语言处理实证方法会议（*EMNLP*）论文集，第1532–1543页。计算语言学协会，多哈，卡塔尔。P. Pu、L. Chen 和 R. Hu。2011年。面向推荐系统的以用户为中心的评估框架。载于《 5*th* *ACM*推荐系统会议论文集，RecSys 2011，第157–164页。M. Quadrana、P. Cremonesi 和 D. Jannach。2018年。序列感知推荐系统。*ACM*计算调查，51(4)。F. Radlinski 和 N. Craswell。2017年。对话式搜索的理论框架。载于*2017*年人类信息交互与检索会议论文集，CHIIR 2017，第117–126页。S. Rendle。2010年。分解机。载于《 10*th* *IEEE*国际数据挖掘会议论文集，ICDM 2010，第995–1000页。

S. Rendle。2012年5月。基于libfm的分解机。*ACM*《智能系统与技术汇刊》，3(3)。S. Rendle、C. Freudenthaler、Z. Gantner 和 L. Schmidt-Thieme。2009年。贝叶斯个性化排序：基于隐式反馈的贝叶斯个性化排序。载于 J. A. Bilmes 和 A. Y. Ng 编，《 25*th*不确定性人工智能会议（*UAI 2009*）论文集，第452–461页。AUAI出版社。S. Rendle、C. Freudenthaler 和 L. Schmidt-Thieme。2010年。用于下一购物篮推荐的个性化马尔可夫链分解。载于《 19*th* 国际万维网会议论文集，WWW 2010，第811–820页。S. Rendle、W. Krichene、L. Zhang 和 J. R. Anderson。2020年。重新审视神经协同过滤与矩阵分解。载于《 14*th* *ACM*推荐系统会议（*RecSys 2020*）论文集，第240–248页。ACM。P. Resnick、N. Iacovou、M. Suchak、P. Bergstrom 和 J. Riedl。1994年。GroupLens：用于网络新闻协同过滤的开放式架构。载于*ACM*计算机支持的协同工作会议论文集，CSCW 1994，第175–186页。F. Ricci、L. Rokach 和 B. Shapira 编。2022年。《推荐系统手册》 3*rd* 版，第3版。施普林格，纽约，纽约州，美国。M. Rossetti、F. Stella 和 M. Zanker。2016年。评估推荐算法时离线结果与在线结果的对比。载于《 10*th* *ACM*推荐系统会议论文集，RecSys 2016，第31–34页。ACM，纽约，纽约州，美国。R. Salakhutdinov 和 A. Mnih。2007年。概率矩阵分解。载于 J. C. Platt、D. Koller、Y. Singer 和 S. T. Roweis 编，第二十一届神经信息处理系统年会论文集（*NIPS 2007*），第1257–1264页。Curran Associates, Inc. M. Sanderson。2010年。基于测试集合的信息检索系统评估。*FoundationsandTrendsinInformationRetrieval*，4(4): 247–375。M. Sanderson 和 W. B. Croft。2012年。信息检索研究的历史。《*IEEE*会议录》，100(百年特刊): 1444–1451。J. Sanz-Cruzado 和 P. Castells。2018年。通过推荐弱关系增强社交网络中的结构多样性。载于 S. Pera、M. D. Ekstrand、X. Amatriain 和 J. O'Donovan 编，《12*th* *ACM*推荐系统会议（*RecSys 2018*）论文集，第233–241页。ACM。B. Sarwar、G. Karypis、J. Konstan 和 J. Riedl。2001。基于项目的协同过滤推荐算法。载于《 10*th* 国际万维网会议论文集，WWW 2001，第285–295页。ACM，纽约，纽约州，美国。J. B. Schafer、J. Konstan 和 J. Riedl。1999年。电子商务中的推荐系统。载于《 1*st* *ACM*电子商务会议论文集，EC 1999，第158–166页。T. Schnabel、A. Swaminathan、A. Singh、N. Chandak 和 T. Joachims。2016年6月。作为试验组的推荐：去偏学习与评估。载于《 3*rd* 国际机器学习会议（*ICML 2016*）论文集，第1670–1679页。《机器学习研究会议录》，谢菲尔德，英国。O. Semerci、A. Gruson、C. Edwards、B. Lacker、C. Gibson 和 V. Radosavljevic。2019年。Spotify主页个性化。载于《 13*th* *ACM*推荐系统会议论文集，RecSys 2019，第527页。

Y. Shi、A. Karatzoglou、L. Baltrunas、M. A. Larson、N. Oliver 和 A. Hanjalic。2012年。CLiMF：通过协同过滤“少即是多”学习最大化倒数排名。载于 P. Cunningham、N. J. Hurley、I. Guy 和 S. S. Anand 编，《 6th *ACM*推荐系统会议论文集，第 139–146 页。ACM。B. Smyth 和 P. McClave。2001。相似性与多样性。载于 《 *4th* 国际基于案例推理会议论文集，ICCBR 2001，第 347–361 页。Springer-Verlag，英国伦敦，英国。B. Song、X. Yang、Y. Cao 和 C. Xu。2018年。神经协同排序。载于《27th *ACM*国际信息与知识管理会议论文集，CIKM 2018，第 1353–1362 页。ACM，美国纽约州纽约市。H. Steck。2010年。在非随机缺失数据上训练和测试推荐系统。载于《 16th *ACM SIGKDD*知识发现与数据挖掘国际会议（*KDD 2010*）论文集，第 713–722 页。ACM，美国纽约州纽约市。H. Steck。2011年。项目流行度与推荐准确率。载于《 5th *ACM*推荐系统会议（*RecSys 2011*）论文集，第 125–132 页。ACM，美国纽约州纽约市。H. Steck。2013。推荐评估：评分预测与排序。载于《 7th *ACM*推荐系统会议（*RecSys 2013*）论文集，第 213–220 页。ACM，美国纽约州纽约市。H. Steck。2018年。校准推荐。载于*ACM RecSys 2018*，第 154–162 页。H. Steck。2019年。适用于稀疏数据的极浅自编码器。载于万维网会议论文集，WWW 2019，第 3251–3257 页。ACM，美国纽约州纽约市。H. Steck、L. Baltrunas、E. Elahi、D. Liang、Y. Raimond 和 J. Basilico。2021。推荐系统的深度学习：Netflix案例研究。《人工智能杂志》，42(3)：7–18。J. Stray。2020年。使人工智能优化与社区福祉保持一致。《国际社区福祉杂志》，3：443–463。F. Sun、J. Liu、J. Wu、C. Pei、X. Lin、W. Ou 和 P. Jiang。2019a。BERT4Rec：基于 Transformer 双向编码器表示的序列推荐。载于《 28th *ACM*国际信息与知识管理会议论文集，CIKM 2019，第 1441–1450 页。ACM，美国纽约州纽约市。F. Sun、J. Liu、J. Wu、C. Pei、X. Lin、W. Ou 和 P. Jiang。2019b。BERT4Rec：基于 Transformer 双向编码器表示的序列推荐。载于《 28th *ACM*国际信息与知识管理会议论文集，CIKM 2019，第 1441–1450 页。ACM，美国纽约州纽约市。https://doi.org/10.1145/3357384.3357895。Z. Sun、Q. Guo、J. Yang、H. Fang、G. Guo、J. Zhang 和 R. Burke。2019c。关于带侧信息推荐的研究评论：综述与研究方向。《电子商务研究与应用》，37：100879。R. S. Sutton 和 A. G. Barto。2021。《强化学习——导论（第二版）》。自适应计算与机器学习。MIT出版社。A. Swaminathan、A. Krishnamurthy、A. Agarwal、M. Dud´ık、J. Langford、D. Jose 和 I. Zitouni。2017年。列表推荐的离策略评估。载于《 31st 神经信息处理系统会议（*NIPS 2017*）论文集，第 3635–3645 页。Curran Associates, Inc.，美国纽约州雷德胡克。

G. Tak´acs 和 D. Tikk。2012年。用于个性化排序的交替最小二乘。载于《 6th *ACM*推荐系统会议（*RecSys 2012*）论文集，第 83–90 页。ACM，美国纽约州纽约市。J. Tang 和 K. Wang。2018年。通过卷积序列嵌入实现个性化 Top-N 序列推荐。载于《 11th *ACM*国际网络搜索与数据挖掘会议论文集，WSDM 2018，第 565–573 页。ACM，美国纽约州纽约市。C. Trattner 和 D. Jannach。2019年。学习根据人类判断推荐相似物品。用户建模与用户适应交互，30：1–49。V. N. Vapnik。1998年。统计学习理论。Wiley-Interscience。S. Vargas 和 P. Castells。2011年。推荐系统新颖性与多样性度量中的排序与相关性。载于《 5th *ACM*推荐系统会议论文集，RecSys 2011，第 109–116 页。ACM，美国纽约州纽约市。S. Vargas 和 P. Castells。2014年。通过向物品推荐用户来提升销售多样性。载于《 8th *ACM*推荐系统会议论文集，RecSys 2014，第 145–152 页。ACM，美国纽约州纽约市。S. Vargas、P. Castells 和 D. Vallet。2011年。推荐系统中的意图导向多样性。载于《 34th 年度国际*ACM SIGIR*信息检索研究与发展会议论文集，SIGIR 2011，第 1211–1212 页。ACM，美国纽约州纽约市。A. Veloso、M. Ribeiro、A. Lacerda、E. Moura、I. Hata 和 N. Ziviani。2014年12月。推荐系统的多目标帕累托高效方法。推荐系统新颖性与多样性专题，*ACM* 信息系统与技术汇刊，5(4)。J. Wasilewski 和 N. Hurley。2016年。使用约束概率潜在语义分析进行意图感知多样化。载于《 10th *ACM*推荐系统会议论文集，RecSys 2016，第 39–42 页。ACM，美国纽约州纽约市。J. Whittaker、S. Looney、A. Reed 和 F. Votta。2021。推荐系统与极端主义内容的放大。《互联网政策评论》，10(2)：1–29。C.-Y. Wu、A. Ahmed、A. Beutel、A. J. Smola 和 H. Jing。2017年。循环推荐网络。载于《 10th *ACM*国际网络搜索与数据挖掘会议论文集，WSDM 2017，第 495–503 页。ACM，美国纽约州纽约市。L. Wu 和 M. Grbovic。2020年。Airbnb 如何告诉你会喜欢巴塞罗那的日落帆船之旅？双边旅游市场中的推荐。载于《 43rd 国际*ACM SIGIR*信息检索研究与发展会议论文集，SIGIR 2020，第 2387–2396 页。ACM，美国纽约州纽约市。L. Wu、X. He、X. Wang、K. Zhang 和 M. Wang。2022年。面向准确率的神经推荐综述：从协同过滤到信息丰富型推荐。*IEEE* 知识与数据工程汇刊，第 1–1 页。数字对象标识符：10.1109/TKDE.2022.3145690。B. Xiao 和 I. Benbasat。2007年。电子商务产品推荐代理：用途、特征与影响。《*MIS* 季刊》，31(1)：137–209。X. Xin、A. Karatzoglou、I. Arapakis 和 J. M. Jose。2020年。面向推荐系统的自监督强化学习，第 931–940 页。SIGIR 2020。ACM，美国纽约州纽约市。

X. Xin、A. Karatzoglou、I. Arapakis 和 J. M. Jose。2022a。面向推荐系统的监督优势演员-评论家。在《 15th *ACM*国际网络搜索与数据挖掘会议论文集，WSDM 2022，页码 1186–1196。ACM，美国纽约州纽约市。X. Xin、T. Pimentel、A. Karatzoglou、P. Ren、K. Christakopoulou 和 Z. Ren。2022b。重新思考用于推荐的强化学习：提示视角。在《 45th国际*ACM SIGIR*信息检索研究与发展会议论文集，SIGIR 2022。S. Yao 和 B. Huang。2017。超越均等性：协同过滤的公平性目标。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，神经信息处理系统进展，第 30 卷。Curran Associates, Inc. Y. Yao 和 F. M. Harper。2018。评判相似性：一项以用户为中心的相关物品推荐研究。在《 12th *ACM*推荐系统会议论文集，RecSys 2018，页码 288–296。F. Yuan、A. Karatzoglou、I. Arapakis、J. M. Jose 和 X. He。2019。用于下一物品推荐的简单卷积生成网络。在《 12th *ACM*国际网络搜索与数据挖掘会议论文集，WSDM 2019，页码 582–590。ACM，美国纽约州纽约市。C. X. Zhai、W. W. Cohen 和 J. Lafferty。2003。超越独立相关性：子主题检索的方法与评估指标。在《 *26th* 年度国际*ACM SIGIR*信息检索研究与发展会议论文集，SIGIR 2003，页码 10–17。ACM，美国纽约州纽约市。J. Zhang、G. Adomavicius、A. Gupta 和 W. Ketter。2020。消费与表现：通过基于智能体的仿真框架理解推荐系统的纵向动态。《信息系统研究》，31(1)。S. Zhang、Y. Tay、L. Yao、A. Sun 和 C. Zhang。2022。推荐系统的深度学习。载于 F. Ricci、L. Rokach 和 B. Shapira 编，《推荐系统手册》，3rd 版，页码 173–210。Springer US，纽约州纽约市。Y. Zhang、F. Feng、X. He、T. Wei、C. Song、G. Ling 和 Y. Zhang。2021。通过因果干预利用推荐中的流行度偏差。在第*44*届国际*ACM SIGIR*信息检索研究与发展会议论文集，SIGIR 2021，页码 11–20。ACM，美国纽约州纽约市。ISBN 9781450380379。C.-N. Ziegler、S. M. McNee、J. A. Konstan 和 G. Lausen。2005。通过主题多样化改进推荐列表。在《 *14th* 国际万维网会议论文集，WWW 2005，页码 22–32。ACM，美国纽约州纽约市。M. Zuckerberg，2018。内容治理与执行蓝图。摘自 Facebook Newsroom 网站：https://www.facebook.com/notes/mark-zuckerberg/ablueprint-for-content-governance-andenforcement/10156443129621634。