1. **Evaluating**
2. **评估方法**

推荐系统的主要评估方法有：离线实验（offline experiments）、用户研究（user studies）和实地测试（field tests）。其中，**离线实验**主要考量推荐系统的性能，比如：预测准确度、覆盖率和流行度等。离线实验的局限性在于，无法直接告诉我们用户对于推荐质量的感知。在离线实验中，值得注意的是：其一，**关注使用的数据及其分布同真实情景中数据间的偏差**。比如，研究人员为了权衡训练成本，可能会过滤数据中的低频用户或项目，这将会在数据中引入新的系统性偏差。一种简单的替代方案是：通过随机抽样用户和项目的方式减少数据量。然而，这种方法可能会导致数据稀疏，使得离线实验更倾向选择擅长处理稀疏数据的算法。至此，可以通过调整数据样本权重的方式，使得稀有样本能够对结果产生更大的影响，从而一定程度上纠正数据稀疏性所带来的偏差。其二，**不同用户评价项目的尺度差异**，也会导致数据本身出现偏差问题。一般而言，可以通过重新抽样（resampling）的方式调整不同用户对项目的评价分布，也可通过重新加权（reweighting）测试数据的方式调整不同用户的评价，从而消除评价尺度不同所带来的偏差问题。其三，**保证用户模型能符合真实情景中的用户**，对于不恰当用户模型的学习，将会迭代训练出模拟表现好但实践表现不佳的推荐模型。

而**用户研究**则在可控的实验环境中，选择若干具有代表性的用户，通过从不同角度测试推荐系统并记录交互行为的方式，评估推荐对用户实际选择的潜在影响，比如：人机交互的感知质量和系统可用性等。然而，这种评估方法的局限性在于：测试用户的代表性值得商榷；代价昂贵，需要权衡测试用户的选取成本和任务数量；测试用户若事先知道测试假设，可能会不自觉地提供支持该假设的证据。**实地测试**则着重于分析推荐系统在真实场景中推荐系统对于用户的影响。迄今为止，在大多数的研究工作中，根据离线实验完成推荐系统的评估更为普遍，用户研究和实地测试相对较少。

1. **显著性检验**

值得注意的是，离线实验关于“某算法的效果优于其它算法”的结果，或许可能是偶然。由此，可以假定一个符合特定情况的零假设，并通过专门的显著性检验实验，用于减少关于“某算法效果更好的结果是偶然”的统计性误差。其中，零假设指代：相较对照或基准算法，新的推荐算法或优化策略没有显著效果、差异或关联。零假设通常提供一个基准，用于辅助评估新的推荐算法是否在统计上显著不同于基准情况。具体而言，在显著性检验的过程中，通常需要将观察到的数据同零假设比较，从而计算出观察到的数据在零假设下出现的概率，即：p-value。至此，若p-value不大于特定的显著性水平，则可以否定预设的零假设并认为观察到的数据具有统计显著性。显而易见的是，更高的显著性水平，拒绝零假设的要求往往也更为严格，亦需要更低的p-value才能得出显著性结论。

**针对观察结果成对的情况**，可以进行**符号检验**，即：假设用户间相互独立，并根据多个用户的配对数据，比较算法A击败算法B的次数。具体而言：首先，设定算法A同算法B效果相同的零假设；其次，计算每个用户分别在算法A和B下的性能度量值；然后，根据前述观察到的配对数据，正差异标记为“+”，负差异标记为“-”，零差异忽略或标记为“0”；再后，统计前述符号序列中“+”和“-”的数量，并使用正态近似等统计方法计算符号检验的统计量；最后，使用统计量和事先设定的显著性水平检验零假设，若算出的p-value小于显著性水平，则拒绝零假设。至此，若两种算法间存在显著差异，还需进一步分析差异原因。若不存在显著差异，则需重新评估或采用其它方法。符号检验，实现简单且无需分布假设，特别适用于小样本的情况。然而，符号检验仅检查了算法A优于B的概率，没有考虑性能差异的具体大小。由此，可以采用**paired Student’s t-test方法**，考虑不同算法性能分数间的平均差异。然而当样本数太少时，此类方法的假设很难验证，且更容易受到异常值的影响。由此，可以考虑使用**Wilcoxon符号秩检验**，此类方法没有对差异进行分布假设。

若无法保证每个测试案例都能被两个算法同时处理，即：**观察结果不成对的情况**，可以使用**Mann-Whitney test方法**，这种方法是Wilcoxon符号秩检验的扩展。具体而言：首先，假设算法A有a个观察结果，算法B有b个观察结果；然后，将两种算法的观察结果汇总，并按照性能度量的大小排序（若并列，则取平均排名）；最后，从排序好的所有结果中随机选取a个结果，若它们的平均排名与来自算法A的a个结果一样好，则说明算法A和B在这些测试案例上的表现相当。这种方法的优点在于：无需事先假设数据服从某种特定的分布，而且即使在小样本的情况下也能给出可靠的结果。

特别的是，**若使用一种性能度量指标，将多种变体算法同基准算法进行比较**，从而获得一个符合特定置信度水平的最佳变体算法。在这种情况下，针对依赖验证集从N+1个算法中筛选出的最佳算法，可进行**Bonferroni校正**，即：若要确保该算法最佳的置信度为1-p，它必须在测试集上以的显著性优于其它N个算法。此外，**若使用多种度量指标，将一种变体算法同基准相比较**。在这种情况下，如若度量指标相互独立，则还可使用Bonferroni校正，否则就需要使用**控制错误发生率**方法了。

1. **Fairness**
2. **基本认知**

过去，关于流行度偏差的研究以及在长尾推荐中提升质量和公平性的努力，被视为推荐系统公平性的早期探索。现在，伴随推荐系统的广泛应用，人们开始愈加关注推荐系统提供的资源及其利益，能否在所影响的不同人或群体间公平分配。换而言之，推荐系统的公平性主要**研究**各种类型的偏差，是如何渗入到推荐系统的数据、算法和输出中去的。

**挑战**在于，推荐系统如何平衡用户的个性化需求、数据的动态特性以及推荐的公平性。同时，尤其在探寻敏感属性对推荐结果公平性影响的研究中，符合要求的数据集并不常见。此外，针对于推荐系统公平性问题，当前尚缺乏统一的公平定义标准以及公平性度量指标，或者说还需要进一步深入理解关于公平性度量标准的要求和行为，从而用于更好地应对偏差本身的多维复杂性和不稳定性。

由此，可行的研究**方向**有：制定通用的去偏差框架，用于度量和解决各种偏差混合影响公平性的问题；深入探究偏差的动态性对于推荐系统公平性的影响；研究人类对于公平推荐的欲望和反应；将公平性推荐研究的概念和方法扩展到更多具体的领域和应用中去。

**1.1 关于偏差**

推荐系统偏差的产生，涉及多方面原因。比如，推荐模型训练所依赖的历史交互数据，本身是混合系统曝光机制和用户自主选择的观测性结果，无法准确且全面的表明用户真实的行为模式。同时，由于流行物品本身能关联更多的用户行为，使得系统更倾向推荐流行物品，从而导致某些物品没有被平等地展示。此外，由于推荐系统本身关于反馈环（feedback loop）的特性，使得用户和推荐系统间存在循环形式的相互干涉，即：推荐系统曝光机制影响用户行为模式的同时，用户由此产生的反馈数据会进一步影响推荐模型训练，从而不断加剧推荐系统的偏差问题。

一般而言，主要存在数据、模型、结果及反馈循环这四种偏差类型。针对于显式数据，主要有**选择和一致性偏差**。前者是因为用户的评分具有主观选择性（比如，用户可能更喜欢对特别好或不好的物品进行评价），使得观测到的评分数据分布不一定能够代表所有评分的分布。后者是因为用户可能会受到其他用户的影响，趋向于给出某种共识性评分，从而无法反映用户自身的真实行为模式。针对于隐式数据，主要有**曝光和位置偏差**。前者是因为只有部分物品曝光给用户，然而其它没有产生过交互的负例物品不一定代表用户不喜欢。后者是因为用户更可能会同推荐列表中位次靠前的物品产生交互，不一定按照用户自身的偏好模式，理性地分配注意力到不同的推荐位次上。针对于推荐模型中的**感应偏差**，主要是因为模型在迭代训练的过程中，所学习的目标函数中没有考虑公平性约束，使得模型过度拟合训练数据中本身存在的偏差。针对于结果中的**流行度偏差和不公平性**。前者是因为极少数的流行物品可能会占据大部分交互记录，在这种物品交互次数呈长尾分布的数据上进行训练，推荐模型将倾向于给流行物品更高的分数。后者是因为对于不平衡数据的学习，可能会导致推荐模型系统性歧视和减少弱势个人或群体的可见性。针对于**反馈循环偏差**，主要是指推荐系统由于自身固有的反馈循环机制，可能会不断加剧推荐的偏差程度。

**1.2 关于公平性**

关于公平性的定义，众说纷纭。从不同的角度出发，公平性大体上可以划分为程序公平（Procedural Fairness）和结果公平（Outcome Fairness）这两方面。其中，**程序公平性**是指决策过程中的程序正义，比如在决策过程中考虑如何更公平地构造和使用输入特征。而**结果公平性**又称为分配式公平（Distributive Fairness），主张在推荐结果中展现公平。具体而言，按照利益相关者的类别来划分，结果公平可细分为**用户公平性**（用户是否获得了公平的推荐）、**项目公平性**（不同项目是否得到了公平对待）及**双边公平性**（同时考虑用户和项目的公平性）。按照资源分配的粒度来划分，又可细分为**群体公平性**（对不同敏感群体的推荐表现平等）和**个体公平性**（相似个体受到相似程度的对待）。此外，还存在旨在消除敏感属性与模型预测间因果关系的**因果公平性**，存在能够灵活过滤任何敏感属性从而允许用户关心期望敏感属性的**个性化公平性**，存在旨在解释模型为何不公平的**可解释性公平性**，存在专注于最大化表现最差个体或群体表现的**最大最小公平性**，存在考虑动态因素对于公平性影响的**动态公平性**。

**1.3 评价指标**

**绝对差异（AD）**衡量劣势群体和优势群体间的效用差异，。其中，群体效用函数用于计算群体的平均评分预测分数或排名性能，即：F1分数或NDCG。若群体间的绝对效用差异AD值较低，则说明推荐系统在这些群体中相对公平。

**方差（Variance）**衡量个体或群体层面的性能分散情况，。其中，，表示全部个体或群体的集合。

**最小-最大差异（MMD）**衡量所有分配效用的最大得分值与最小得分值间的差异，可以反映多个项目组曝光机会的差异，。

**熵（Entrop）**通常用于表明系统的无序程度，可以用来评估推荐过程中项目曝光机会的不平等，。

**KL-散度（KL-Divergence）**衡量概率分布间的差异，可用于计算项目组的曝光分布与它们在推荐中的历史曝光间的差异，。其中，较低的KL-散度值表明推荐更公平。

**1.4 公平推荐方法**

公平推荐方法，大体上可以划分为：预处理、处理中和后处理三个大类。

关于**预处理**。推荐系统在数据公平性方面的研究有限，当前更多借鉴机器学习文献中的相关技术。比如，数据集中某些个体或群体的样本太少，或许会影响针对他们的推荐质量，因此可以采用数据增强的方法收集更多的特定标记数据，从而缓解针对某些弱势个体或群体的推荐偏差。或者，也可通过数据重采样方法来平衡数据分布，使得敏感个体或群体的样本数据量接近。此外，还可扰动样本不敏感特征的值，使得不敏感特征取值及分布不再同特定敏感属性相关联，从而一定程度上缓解公平性问题。然而，当前的数据公平性方法，尚未充分考虑用户-物品交互数据的动态性，在推荐系统领域的适用性有限。

关于**处理中**。在构建公平性推荐系统的过程中，实际上需要处理涉及公平性和推荐性能等方面的多目标优化问题。一般而言，可以在目标函数中引入考虑公平性问题的正则化项，从而最小化推荐模型关于敏感属性的预测误差。也可通过约束优化的方式，控制用户对个别物品的关注程度。同时，还可以使用强化学习方法，针对于代理程序设立公平性阈值约束，从而确保在公平策略学习的过程中考虑到偏差问题。此外，对抗学习或因果图等方法，也可在过程中缓解偏差问题。

关于**后处理**。实际上旨在通过二次处理推荐结果的方式，提高最终推荐的公平性。比如，可以使用逐槽、用户级或全局级重排名等方法，重新组织推荐列表中的项目次序，从而达到平衡个性化推荐性能和公平性的目的。

值得注意的是，针对推荐系统的进一步公平性约束，降低系统推荐质量的同时，不一定能够提高特定个体或群体的推荐效用。此外，从工程实现的角度出发，通过数据或算法介入并不总是实现公平的最佳途径，更重要的是要理解为什么一些用户的体验质量较差，并采取针对性措施改善他们的体验。这种思路不仅可以提高推荐系统的整体公平性，还能增加推荐系统的用户满意度和整体效能。

1. **服务提供者公平性（Provider Fairness）**

若从不同涉众的角度出发，讨论推荐系统公平性问题，消费者公平性主要关注向消费者提供推荐的质量，服务提供者公平性则主要**关注**推荐对提供者的效用。其中，相较于消费者公平性仅需要通过查看单次推荐结果，即可评估对应用户的推荐质量；提供者公平性**必须**在一系列推荐结果被提供给用户后，才能进行效用的累积测量。此外，虽然增强推荐的多样性或许能促进提供者公平，但是推荐多样性和提供者公平性在动机上存在本质的区别。比如，推荐多样性主要集中在消费者的意图上，是满足用户广泛主题需求的结果。提供者公平性则出于正义考量的**动机**，致力于服务提供者们都能获得相对公平的机会，使其服务内容或产品被适度曝光发现。

具体而言，服务提供者公平性将推荐机会视为一种资源，并重点关注推荐对服务提供者的效用或曝光度（给特定项目（组）提供推荐机会的价值）。针对**效用的计算方式**，即可以简单假设推荐列表中所有项目都有相同的机会被用户点击，因此给每个项目分配相同的固定效用值（忽略了用户对不同项目的偏好和关注程度）。也可以根据项目在推荐列表中的位置进行加权处理，从而完成额外考虑用户关注程度的加权效用计算。

更进一步，服务提供者公平性，又可细分为提供者的个体或群体公平性。其中，**提供者个体公平性**，要求相似的服务提供者应该从推荐系统中获得类似的效用，可以根据提供者与用户的相关性来定义“相似”，即：如果某用户同时喜欢两件物品，那么这两件物品相似。此外，也可以利用潜在因素空间的概念，将相似的物品聚类在一起，并在向用户推荐物品的过程中保留这种相似性。**提供者群体公平性**，则要求在保证个性化推荐质量的前提下，确保不同群体的服务提供者受到公平对待。提供者群体公平性的实现方式比较多样，比如：可以强制推荐与敏感属性间的统计独立性，即：若某物品出现在特定推荐列表中的概率和其敏感属性无关，则认定推荐是公平的；可以使用曝光度对多个推荐列表进行评估，从而衡量每种服务提供者群体是否接收到适当水平的曝光；也可以测量推荐列表内服务提供者群体分布同目标分布间的差异，来衡量受保护群体在推荐中的公平代表程度。