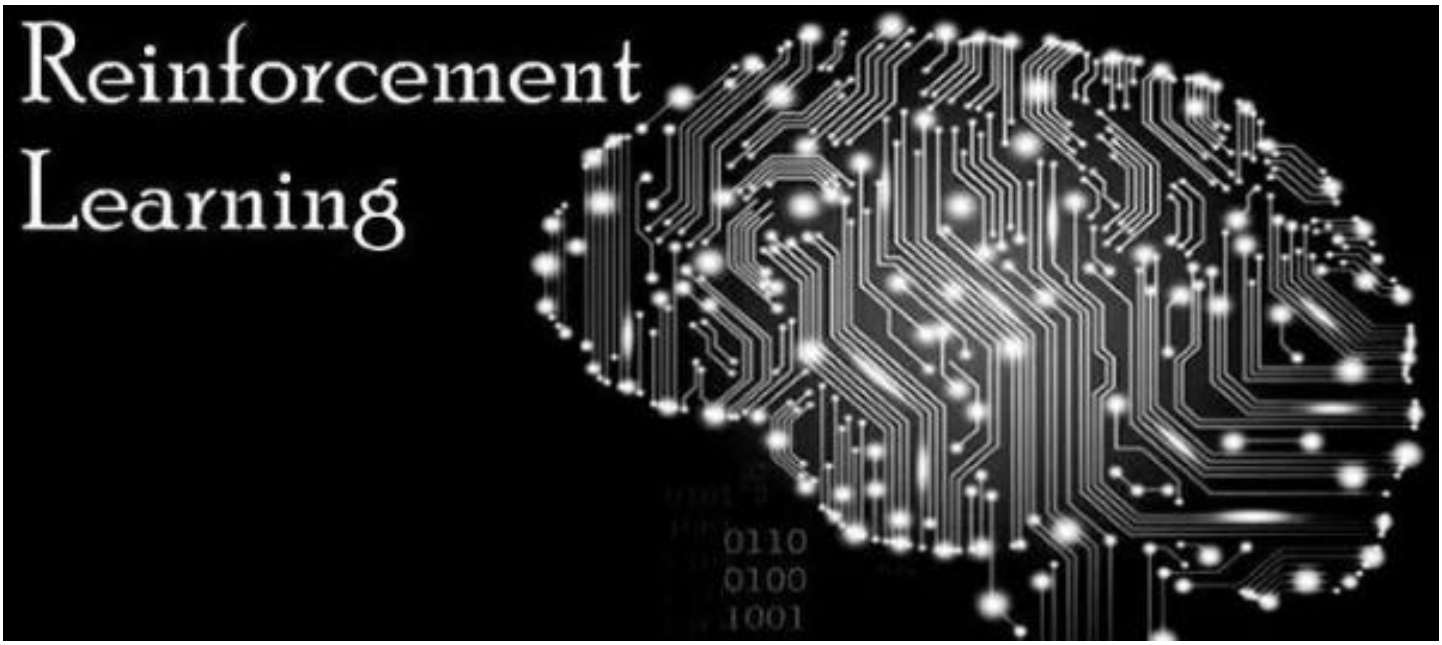
知乎



干货|个性化推荐系统五大研究热点之强化学习（三）

第四范式...
已认证的官方帐号

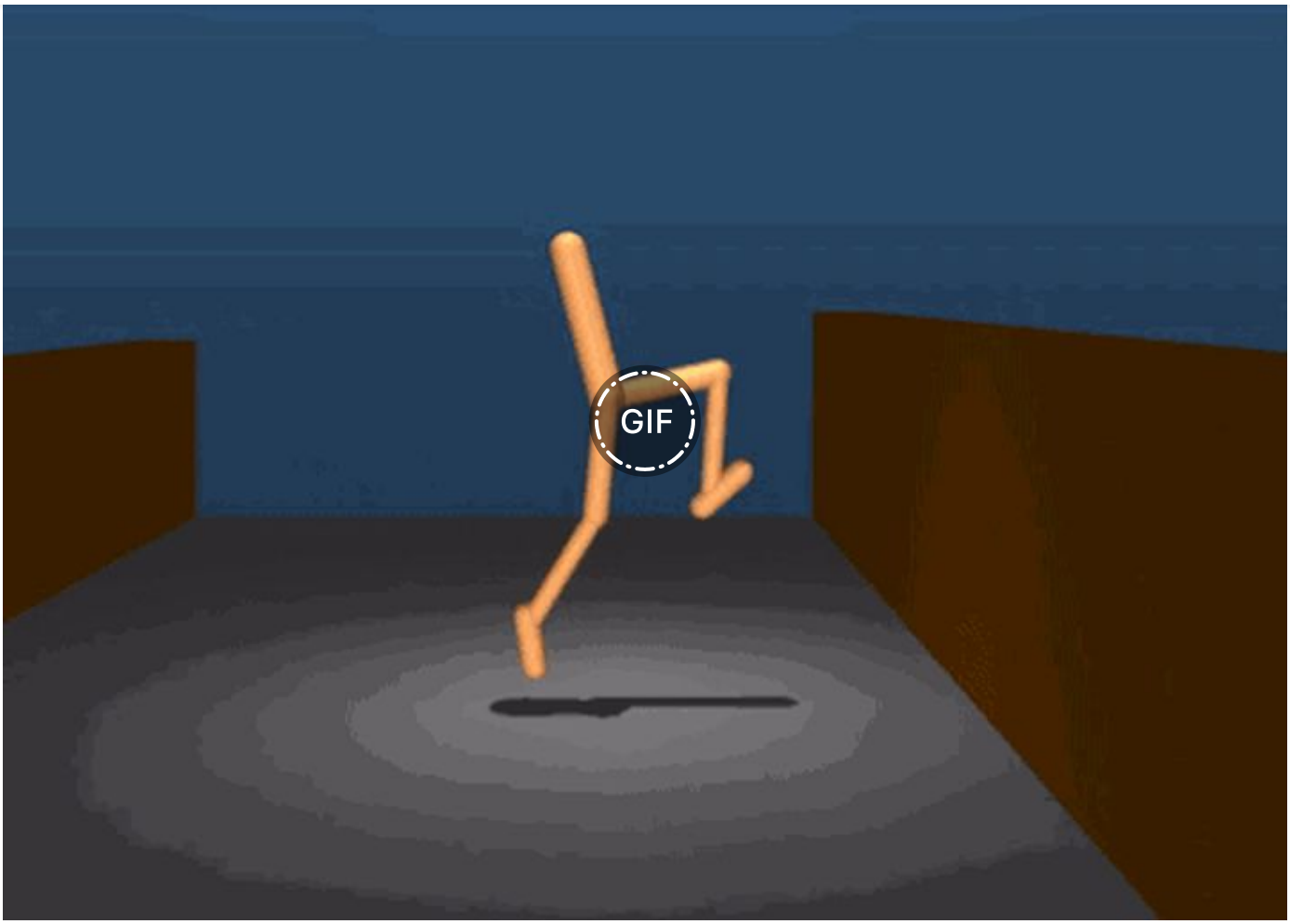
关注他

21 人赞同了该文章

> 【编者按】微软亚洲研究院社会计算组的研究员们从**深度学习、知识图谱、强化学习、用户画像、可解释性推荐**等五个方面，展望了未来推荐系统发展的方向。
>
> 在前两篇文章中，我们分别介绍了**深度学习**技术和**知识图谱**在推荐系统中的应用以及未来可能的研究方向。在今天的文章中，我们将介绍强化学习在推荐系统中的应用。

通过融合深度学习与知识图谱技术，推荐系统的性能取得了大幅的提升。然而，多数的推荐系统仍是以一步到位的方式建立的：**它们有着类似的搭建方式，即在充分获取用户历史数据的前提下，设计并训练特定的监督模型，从而得到用户对于不同物品的喜好程度。这些训练好的模型在部署上线后可以为特定用户识别出最具吸引力的物品，为其做出个性化推荐。**在此，人们往往假设用户数据已充分获取，且其行为会在较长时间之内保持稳定，使得上述过程中所建立的推荐模型得以应付实际中的需求。

知乎



然而对于诸多现实场景，例如电子商务或者在线新闻平台，用户与推荐系统之间往往会发生持续密切的交互行为。**在这一过程中，用户的反馈将弥补可能的数据缺失，同时有力地揭示其当前的行为特征，从而为系统进行更加精准的个性化推荐提供重要的依据。**

强化学习为解决这个问题提供了有力支持。依照用户的行为特征，我们将涉及到的推荐场景划分为静态与动态，并分别对其进行讨论。

1. 静态场景下的强化推荐

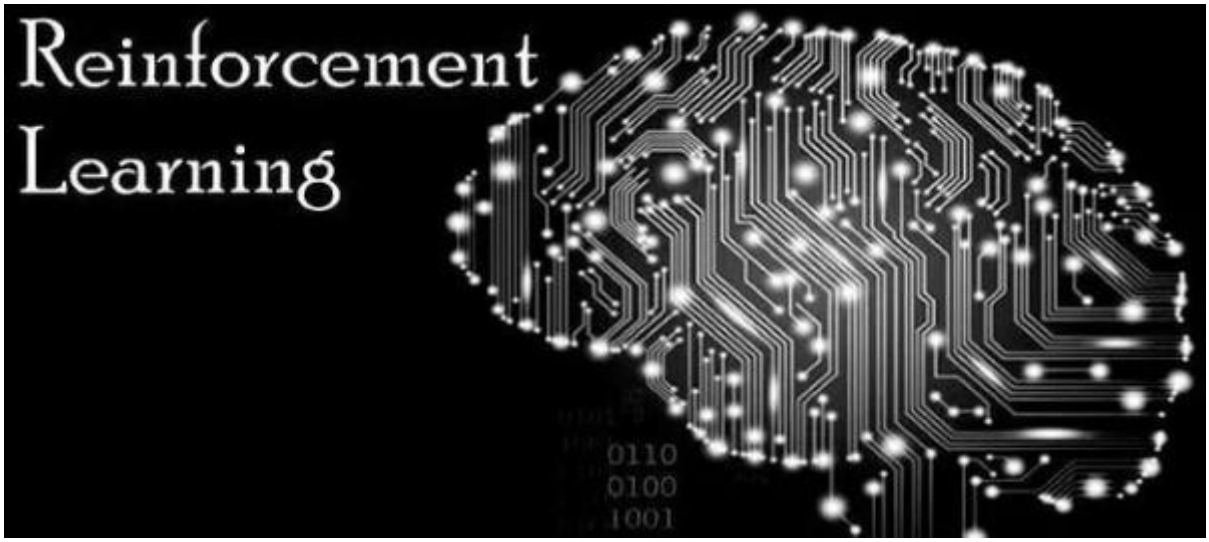
在静态场景之下，用户的行为特征在与系统的交互过程中保持稳定不变。对于这一场景，一类有代表性的工作是**基于上下文多臂老虎机（contextual multi-armed bandit）的推荐系统**，它的发展为克服推荐场景中的冷启动问题提供了行之有效的解决方案。

在许多现实应用中，用户的历史行为往往服从特定的长尾分布，即大多数用户仅仅产生规模有限的历史数据，而极少的用户则会生成较为充足的历史数据。这一现象所带来的数据稀疏问题使得传统模型在很多时候难以得到令人满意的实际效果。

引发极大的探索开销，使得它在现实中并不具备可行性。

为使主动式探索具备可行的效用开销，人们尝试借助多臂老虎机问题所带来的启发。多臂老虎机问题旨在于“探索-利用”间做出最优的权衡，为此诸多经典算法被相继提出。尽管不同的算法有着不同的实施机制，它们的设计都本着一个共同的原则。

具体说来，系统在做出推荐的时候会综合考虑物品的推荐效用以及累积尝试。较高的推荐效用预示着较低的探索开销，而较低的累积尝试则表明较高的不确定性。为此，不同的算法都会设计特定的整合机制，使得同时具备较高推荐效用与不确定性物品可以得到优先尝试。

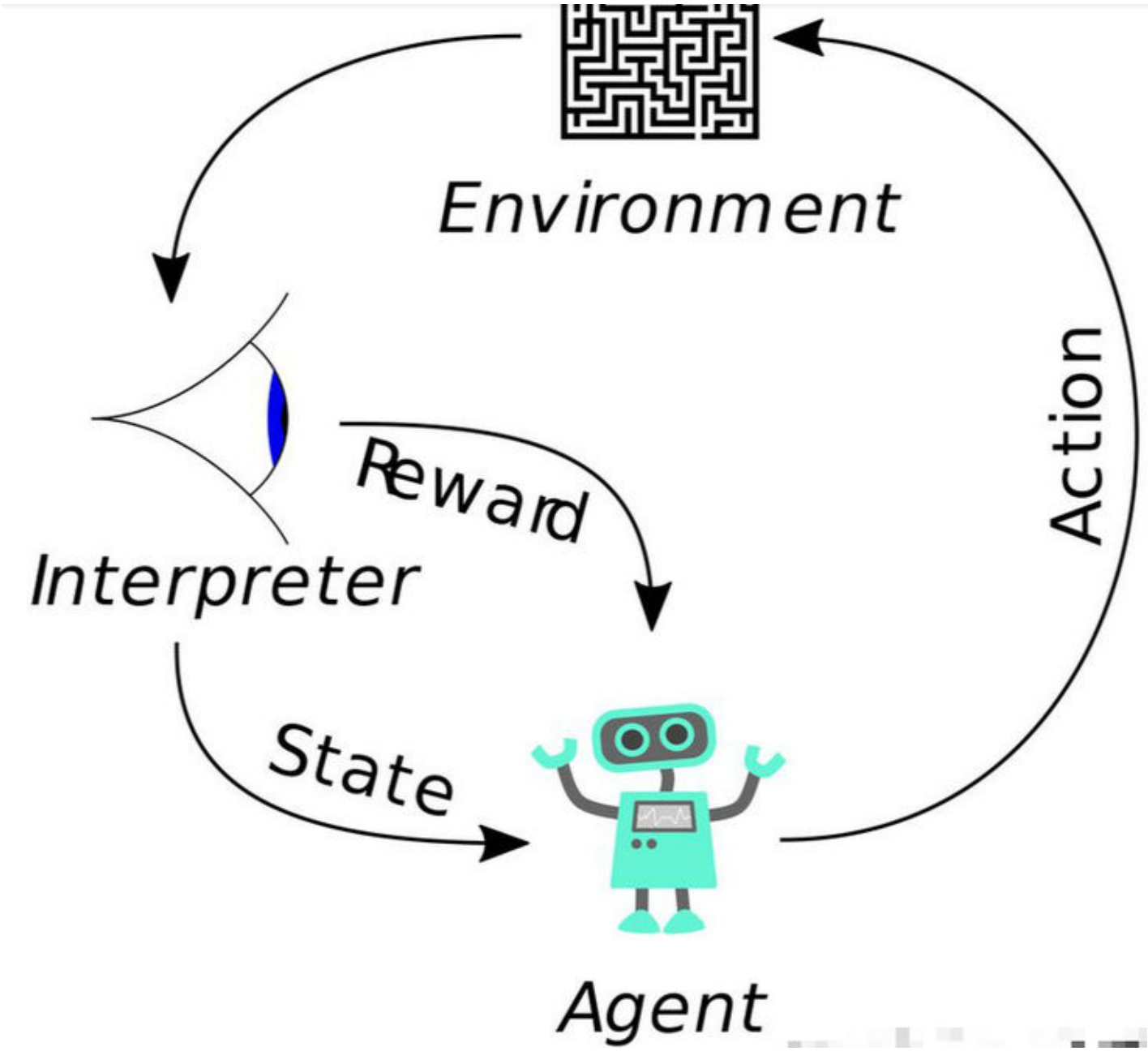


2. 动态场景下的强化推荐

在多臂老虎机的设定场景下，用户的实时特征被假设为固定不变的，因此算法并未涉及用户行为发生动态迁移的情况。然而对于诸多现实中的推荐场景，用户行为往往会在交互过程中不断变化。**这就要求推荐系统依照用户反馈精确估计其状态发展，并为之制定优化的推荐策略。**

具体来讲，一个理想的推荐系统应满足如下双方面的属性。一方面，推荐决策需要充分基于用户过往的反馈数据；另一方面，推荐系统需要优化整个交互过程之中的全局收益。强化学习为实现上述目标提供了有力的技术支持。

知乎



在强化学习的框架之下，推荐系统被视作一个**智能体（agent）**，用户当前的行为特征被抽象成为**状态（state）**，待推荐的对象（如候选新闻）则被当作**动作（action）**。在每次推荐交互中，系统依据用户的状态，选择合适的动作，以最大化特定的长效目标（如点击总数或停留时长）。推荐系统与用户交互过程中所产生的行为数据被组织成为**经验（experience）**，用以记录相应动作产生的**奖励（reward）**以及**状态转移（state-transition）**。基于不断积累的经验，强化学习算法得出**策略（policy）**，用以指导特定状态下最优的动作选取。

我们近期将强化学习成功应用于**必应个性化新闻推荐**（***DRN: A Deep Reinforcement Learning Framework for News Recommendation*, WWW 2018**）。得益于算法的序列化决策能力及其对长效目标的优化，强化学习必将服务于更为广泛的现实场景，从而极大地改善推荐系统的用户感知与个性化能力。

知乎

强化学习推荐算法尚有诸多富有挑战性的问题亟待解决。

现行主流的深度强化学习算法都试图避开对环境的建模，而直接进行策略学习（即model-free）。这就要求海量的经验数据以获取最优的推荐策略。然而，推荐场景下的可获取的交互数据往往规模有限且奖励信号稀疏（reward-sparsity），这就使得简单地套用既有算法难以取得令人满意的实际效果。**如何运用有限的用户交互得到有效的决策模型将是算法进一步提升的主要方向。**

此外，现实中人们往往需要对不同推荐场景进行独立的策略学习。不同场景下的策略互不相同，这就使得人们不得不花费大量精力以对每个场景都进行充分的数据采集。同时，由于不具备通用性，既有策略难以迅速适应新的推荐场景。面对这些挑战，**人们需要尽可能地提出通用策略的学习机制，以打通算法在不同推荐场景间的壁垒，并增强其在变化场景中的鲁棒性（robustness）。**

下一篇文章我们将围绕“**推荐系统中的用户画像**”的研究展开讨论。想要了解关于推荐系统的更多研究热点，还请持续关注。

相关阅读：

干货|个性化推荐系统五大研究热点之知识图谱（二）

干货 | 个性化推荐系统五大研究热点之深度学习（一）

AutoML在推荐系统中的应用

搭建推荐系统快速入门，只需五步！

欢迎大家点赞、收藏，将更多技术知识分享给身边的好友——你的认可就是我们努力的方向。

本账号为第四范式智能推荐产品先荐的官方知乎账号。本账号立足于计算机领域，特别是人工智能相关的前沿研究，旨在把更多与人工智能相关的知识分享给公众，从专业的角度促进公众对人工智能的理解；同时也希望为人工智能相关人员提供一个讨论、交流、学习的开放平台，从而早日让每个人都享受到人工智能创造的价值。

第四范式每一位成员都为人工智能落地贡献了自己的力量，在这个账号下你可以阅读来自计算机领域的学术前沿、知识干货、行业资讯等。