**阿里妈妈新一代智能检索模型亮相WWW2018**



4月25日，阿里妈妈搜索直通车算法团队携最新论文亮相WWW 2018（The International World Wide Web Conference）。今年的WWW大会在法国里昂召开，阿里妈妈高级算法工程师闫肃作为该论文第一作者在大会现场进行了口头报告。

阿里妈妈入选此次大会的论文名为《Beyond Keywords and Relevance: A Personalized Ad Retrieval Framework in E-Commerce Sponsored Search》，评委一致认为该方法是对传统搜索广告检索框架的重新定义，实际上，这也是搜索直通车首次公开其自研的新一代智能检索模型。



阿里搜索直通车广告业务有着巨大的体量和影响力，其技术工作有着非常高的挑战性。面对淘系搜索广告业务场景中真实存在的各种痛点和挑战，阿里技术一线的同学们不断地进行技术探索，通过一次次的技术突破和创新，解决了大量的业务难题。本次阿里妈妈在WWW 2018公开的新一代智能检索模型工作，就是搜索直通车算法团队的同学一次从实践出发，将技术创新和业务诉求相结合的范例。

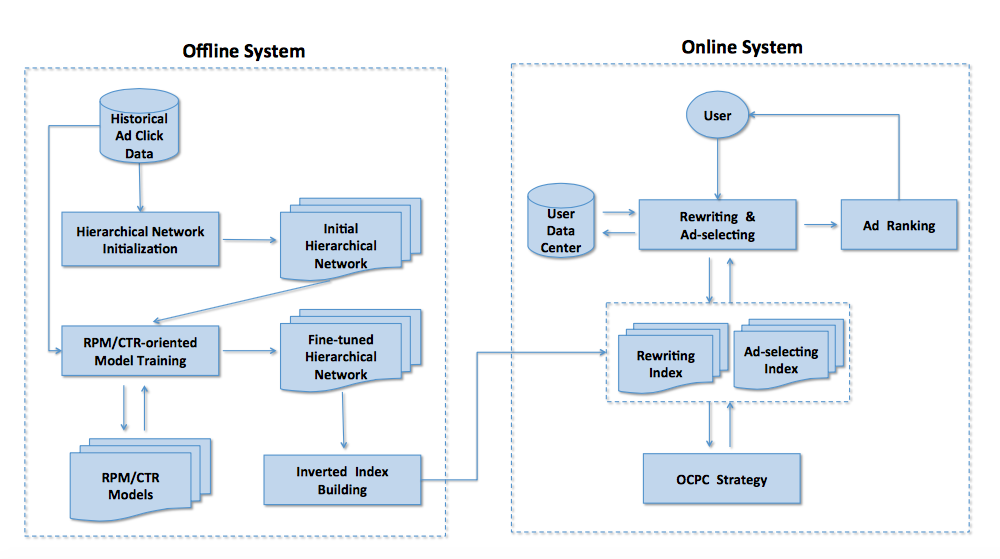


图1：“新一代”搜索广告智能检索框架

在论文中，阿里的技术同学突破了以“关键词”和“相关性”为核心的传统搜索广告检索框架，提出了新一代的搜索广告智能检索模型。新一代搜索广告智能检索模型引入用户行为异构图挖掘、机器学习等相关技术，通过模型学习的方式智能构建索引，解决了传统搜索广告检索系统不能解决的种种痛点，在搜索直通车业务线上取得了出色的效果，给广告商、用户和平台带来了三赢。

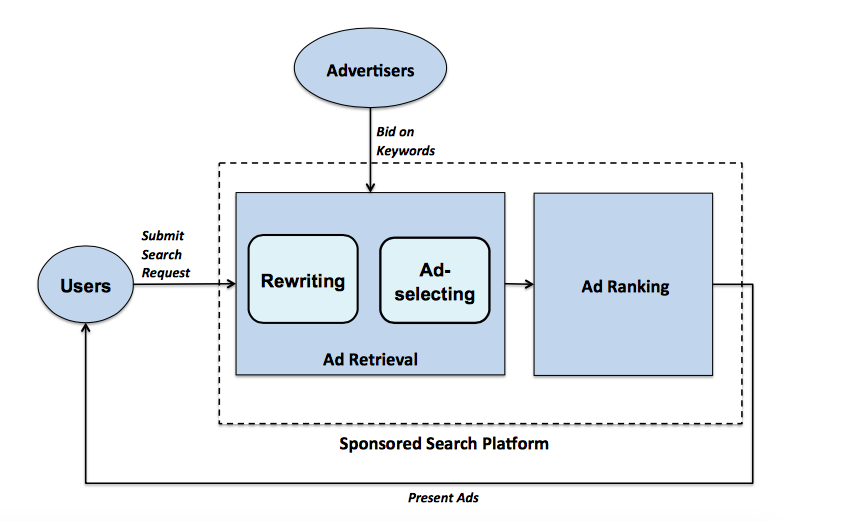


图2：搜索广告系统由三方参与：广告商、用户和系统平台

在搜索广告系统中，每一次搜索广告的展示、点击和转化都需要三个参与方（广告商、用户和平台）的密切合作。平台是用户搜索请求和广告商投放的广告之间的桥梁，进行着流量匹配、广告展现等工作。其中，广告检索模块负责理解用户的搜索意图，快速准确地从海量广告中检索出一个小规模的高质量广告候选集。广告检索模块需要兼顾系统的效果与效率，因此在算法工作中存在着巨大的技术挑战。

在传统的搜索广告系统中，广告商必须为自己的广告选择竞价关键词。平台进行广告检索时会受到竞价关键词的约束。如果广告商没有事先为广告购买相应的关键词，那么即使用户搜索请求与广告紧密相关，平台也不会检索回这些广告。但是，受限于市场信息的缺失和投放管理的巨大成本，广告商有时并不能及时准确地为自己的广告选择出最合适的关键词。在这种情况下，广告检索算法不能实现最优的流量匹配，给广告商、用户和平台三方均带来了损失。

此外，传统的搜索广告检索模型只关注于搜索请求与广告之间的相关性；这往往和平台的目标（RPM、CTR、GMV等）并不完全一致。如何在考虑相关性的同时，兼顾平台目标和用户体验，是广告检索模型需要解决的巨大难点。

近年来，越来越多的个性化信息被引入电商搜索广告系统，如用户在平台上的浏览、点击、交易等行为。一方面，这些个性化信息能够帮助广告检索模型更好地理解用户的搜索意图。但另一方面，个性化信息也给广告检索带了新的挑战：面对从各种复杂丰富的个性化信号通道检索回的广告，检索模型需要能够高效、准确地对其按照统一标准快速排序。这个问题，在目前已知的工作中，均没有得到有效地解决。

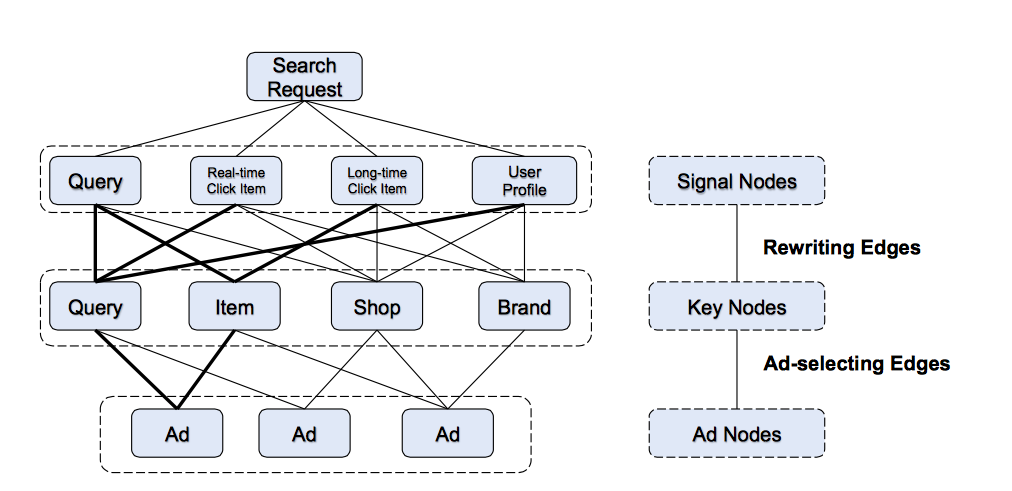
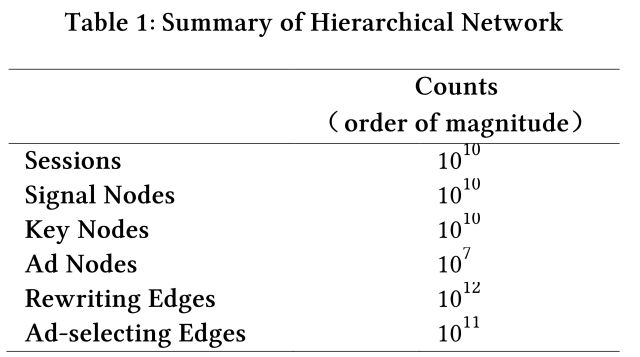
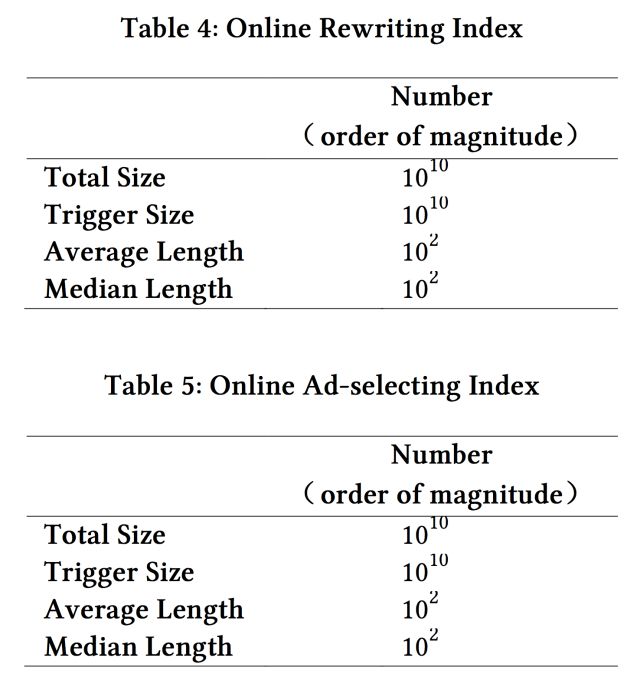


图3：用户行为异构图图示例。图中包含了三种节点：用户搜索信号、广告检索键和广告。用户搜索信号和广告检索键之间的边表示改写，广告检索键和广告之间的边表示广告海选。

面对上述传统搜索广告检索系统中存在的各种难题和挑战，阿里妈妈搜索直通车算法团队的同学提出了一种创新的搜索广告智能检索系统。新的智能检索系统首先使用用户在平台上的历史行为构建出一张庞大复杂的用户行为异构图。异构图中节点分别表示“用户搜索信号”、“广告检索键”和“广告”，边分别表示“用户搜索意图信号改写”关系和“广告召回”关系。接着，检索系统面向平台RPM、CTR等指标，学习异构图中边的权重，挖掘出重要的改写关系和广告召回关系。

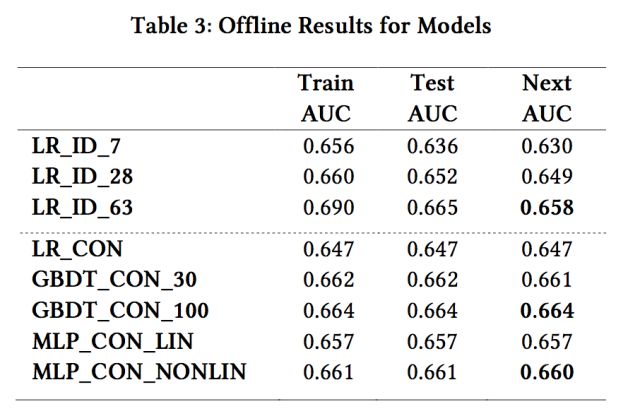
这样，通过对异构图的深入挖掘，检索系统同时进行了“用户搜索意图信号改写”和“广告召回”两个检索子任务的统一联合学习。最后，检索系统根据模型的边挖掘结果，自动构建相应的“改写索引”和“广告召回索引”。通过两个模型智能构建的索引，检索系统将用户行为异构图和模型挖掘结果存储下来，实现了对线上搜索请求的高效检索。由于新的智能检索模型不再强制要求广告商购买关键词，所以新的检索系统使用OCPC策略，在保证广告商ROI的基础上，决定广告的点击收费。

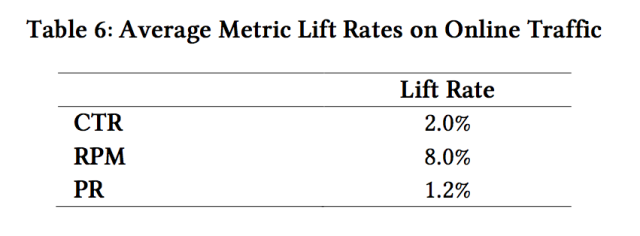




阿里妈妈搜索直通车业务有着巨大的体量和规模庞大的用户数据，因此新的智能广告检索模型在实际落地过程中，也面临着各种技术挑战。例如，在新的广告检索系统中，用户行为异构图庞大复杂，包含了上百亿的节点和上万亿的边，使得模型训练非常困难。为了兼顾检索系统的的效果和性能，阿里技术同学提出了多种异构图初始化方法，在尽量保留重要关系的前提下，实现了对异构图的剪枝，给模型的训练提供了良好的起点。

又例如，在搜索广告检索阶段，为了提高检索效率，模型无法获取足够多的信息或者使用过于复杂的特征。因此，在新的检索模型中，阿里技术同学有针对性地设计了两种“粒度”不同特征：稀疏特征和连续特征。前者是一种细粒度的特征，保证了模型效果；后者则是一种粗粒度特征，用于提高模型的覆盖能力和稳定性。





新的智能搜索广告检索模型，在搜索直通车平台上取得了出色的效果，给广告商、用户和平台带来了三赢：新的检索模型通过OCPC的方式自动为广告出价，在保证了广告商的ROI前提下，把广告商从繁重的买词任务中解放了出来；通过引入丰富的个性化信号，新的检索模型能够更好地理解用户的搜索意图，达成更准确的流量匹配，提升了用户体验；新的检索模型不再单纯以相关性为目标，而是综合考虑平台的目标和用户的体验，提升了平台收益，也维护了平台的生态环境。

阿里妈妈在WWW 2018论文中公布的新一代搜索广告智能检索模型，不仅仅是国际一流的学术成果，更是搜索直通车算法团队的一线技术同学，以技术为驱动，服务广大淘宝、天猫用户和卖家的真实实践。

值得一提的是，WWW大会作为全球互联网探讨未来发展的首要国际学术会议，聚集来自国际著名大学、研究机构、跨国企业和国际标准化组织的一流学者和工业界精英，对论文的要求十分严苛，近几年的论文录取率约为15%左右。

随着万维网的不断发展，WWW大会从最开始的注重万维网的系统部署与架构分析，发展到关注超链接和多媒体，再到如今集中于人工智能、深度学习、安全与隐私保护、内容分析、推荐与挖掘等领域，越来越吸引阿里巴巴、谷歌、腾讯等互联网巨头的加入，推动理论前沿和工业场景应用的紧密结合，持续为各国信息化建设提供了重要的技术支持。

论文原文链接：https://arxiv.org/abs/1712.10110