调研报告: 推荐系统与机器学习

方盛俊 (201300035、318483724@qq.com)

(南京大学 人工智能系, 南京 210093)

摘 要: 由于当下数据爆炸性增长与人们选择日渐多元化, 人们的对物品的需求呈现出长尾效应, 各大平台都急需 “千人千面” 式的针对每一个人独特爱好的商品推荐手段, 也就是推荐系统. 传统的推荐系统算法有基于内容的推荐算法和基于协同的推荐算法, 但是这些传统的算法有着其自己的缺点, 例如过于依赖历史数据和数据量大时性能变差等缺点. 近几年迅速发展的机器学习尤其是深度学习, 能够在一定程度上弥补这些不足, 并在推荐精度上取得更好的效果. 这些推荐系统算法被广泛用于国内外的各大互联网公司中, 例如谷歌, 亚马逊, 淘宝, 字节跳动等.

关键词: 推荐系统;机器学习;深度学习;协同推荐

中图法分类号: TP301 文献标识码: A

1 行业背景

1.1 数据增长与长尾效应

随着互联网行业的发展, 我们正式进入了一个信息爆炸的时代, 具体表现为: 不同种类的商品繁多, 新闻信息的极具增长, 人人都是新闻来源的自媒体模式, 各大图文与视频多媒体平台出现. 人们在多种多样的选择面前眼花缭乱, 因此广告信息也铺天盖地, 各大平台也出现了意见领袖和直播带货的模式. 而由于人们个人兴趣的多样性, 导致了商品销售呈现了长尾效应, 冷门商品的需求量小, 但也不会接近于零, 并且冷门商品的销售总和依然能够达到一个很可观的数额. 以上种种, 都表明了当前的互联网平台都急需个性化的推荐技术.

1.2 推荐系统

推荐系统, 就是依据用户的偏好推荐其最有可能感兴趣的内容. 以前的新闻平台往往采用的是门户网站的形式, 例如雅虎和新浪新闻, 是中心化加编辑推荐的方式进行新闻分发; 如今的新型新闻平台, 例如字节跳动的今日头条, 采用的是使用推荐系统基于用户偏好推送的个性化新闻. 再例如, 如今的电商平台, 例如亚马逊和淘宝, 均拥有 “猜你喜欢” 模块, 基于推荐系统和用户过往的浏览数据购物数据进行个性化的商品推荐. 而当今的一些 UGC (用户生产内容) 平台也纷纷向基于推荐系统的个性化内容推荐发展, 例如抖音, 知乎, Bilibili 这类视频平台或问答平台.

甚至可以说, 基于机器学习尤其是深度学习大发展基础上的推荐系统的发展, 正是近十年互联网行业发展的一个大方向, 任何一个有着庞大用户量的互联网的平台均在推荐系统上投入了巨量的资源和人力.

2 关键技术

2.1 基于内容标签的推荐算法

基于内容标签的算法是最为传统的，我们可以在各大内容网站看到它们的身影，例如商品的类别标签，新闻的分类标签，音乐的流派标签等，然后结合用户的历史行为，便能进行简单而有效的推荐．例如，我们将一部电影根据分为爱情和科幻两个类别，然后通过用户的评分数据得出用户对爱情和科幻类电影的偏好，如果用户更喜欢科幻电影，就给他推荐更多的科幻电影．

但是这种推荐算法严重依赖于物品的内容标签之类的数据，如果依靠人工标注，不仅工作量大，还不容易保证准确律．当前，我们也发展出了一些自动化提取标签的方法，比如 TF-IDF 算法．

2.2 基于协同过滤的推荐算法

协同过滤算法是目前最为主流的推荐算法，该算法最早由亚马逊提出与应用．什么是协同过滤算法？简单来说，就是如果甲和乙都购买过物品 A，而乙也买过物品 B，那么我们可以合理推断甲也很有可能会想要购买物品 B．基于这种朴素而有效的想法，现代发展出了很多种协同过滤算法，主要分为三种，基于用户的协同过滤算法，基于物品的协同过滤算法和基于模型的协同过滤算法．

前两者我们很容易理解其中的思想，基于用户的协同过滤是想要找出与你最相近的用户， 基于物品的协同过滤算法是想要找出和你消费过的物品最相近的物品，但是基于模型的协同过滤算法又是什么？我们知道，前两者虽然应用广泛，但是过于依赖历史数据，数据稀疏时精确度会显著下载，这便是长尾效应．所以我们需要基于模型的协同过滤算法，包括聚类模型，贝叶斯网络和奇异值分解等．

除了这些常用的推荐算法，我们还要考虑许多其他因素，例如上下文和用户画像．上下文指当前的地理，环境和时间因素等．人们在夏天的时候更希望看见冰凉的事物，南方人会喜欢水乡相关的旅游视频，诸如此类．而用户画像更为重要，用户的性别与年龄，兴趣爱好，比如 18 岁男生一般不会浏览母婴相关的内容．

2.3 推荐算法具体步骤

推荐系统的过程一般分为几步，召回，粗排，精排，混排．

召回过程是在数以万亿记的内容库中，去除绝大部分不相关的内容的过程，即将一个稀疏矩阵转变为一个稠密矩阵的过程．这个过程中一个非常重要的步骤是特征工程，其将原始数据转化为更有代表性的数据．常见的例子是，假如我们要判断一个人胖还是不胖，我们不会仅要看体重数据，也要看身高数据．但是如果给了一个人具体的身高和体重，我们并不能很直接地判断出这个人是否肥胖．但是通过引入 BMI 指数, BMI=体重/(身高^2)．我们便能非常清晰地看出这个人身材如何，这就是一个非常简单的特征工程例子．

粗排，让我们能够从召回的内容集合中对内容进行粗略地排序．此时我们并不能像召回步骤那样直接去除低

相关性的内容，因为直接去掉低相关性的内容会影响多方面的用户体验．首先是召回律，如果将大量内容去除，会降低用户找到相关内容的机会，影响到召回率，甚至会降低准确律，还有新颖性和多样性会因此显著降低．我们不应该直接去除这些内容，而是对其进行排序，减少其出现可能性，而不是让他们完全消失．

精排，对于一部分粗排排名很高的的内容，我们应该对其进行精排，让其更符合用户的需求．例如搜索引擎第一页的内容，我们应该把相关性最高的官网，百科，问答放置在第一位，将仅仅只是提及的网页置于后面．

混排，对于高准确性的内容，例如用户在抖音上刷了很多撸猫相关视频之后，用户也很容易审美疲劳，想看一点其他内容．这时候我们就要考虑新颖性，多样性，惊喜性等相关因素．这时候我们就需要混排，随机将一部分低相关度的内容提到前面，让用户有机会浏览．

还有一些其他的相关技术，例如冷启动问题．一个新用户，新内容加入推荐系统时，系统只有很少于其相关的信息，甚至没有与其相关的信息，这被称为冷启动问题．我们要考虑相关的解决方案，例如推荐被绝大多数人喜欢的内容，在用户注册时让其选择相关的标签等方式．

2.4 推荐算法与机器学习

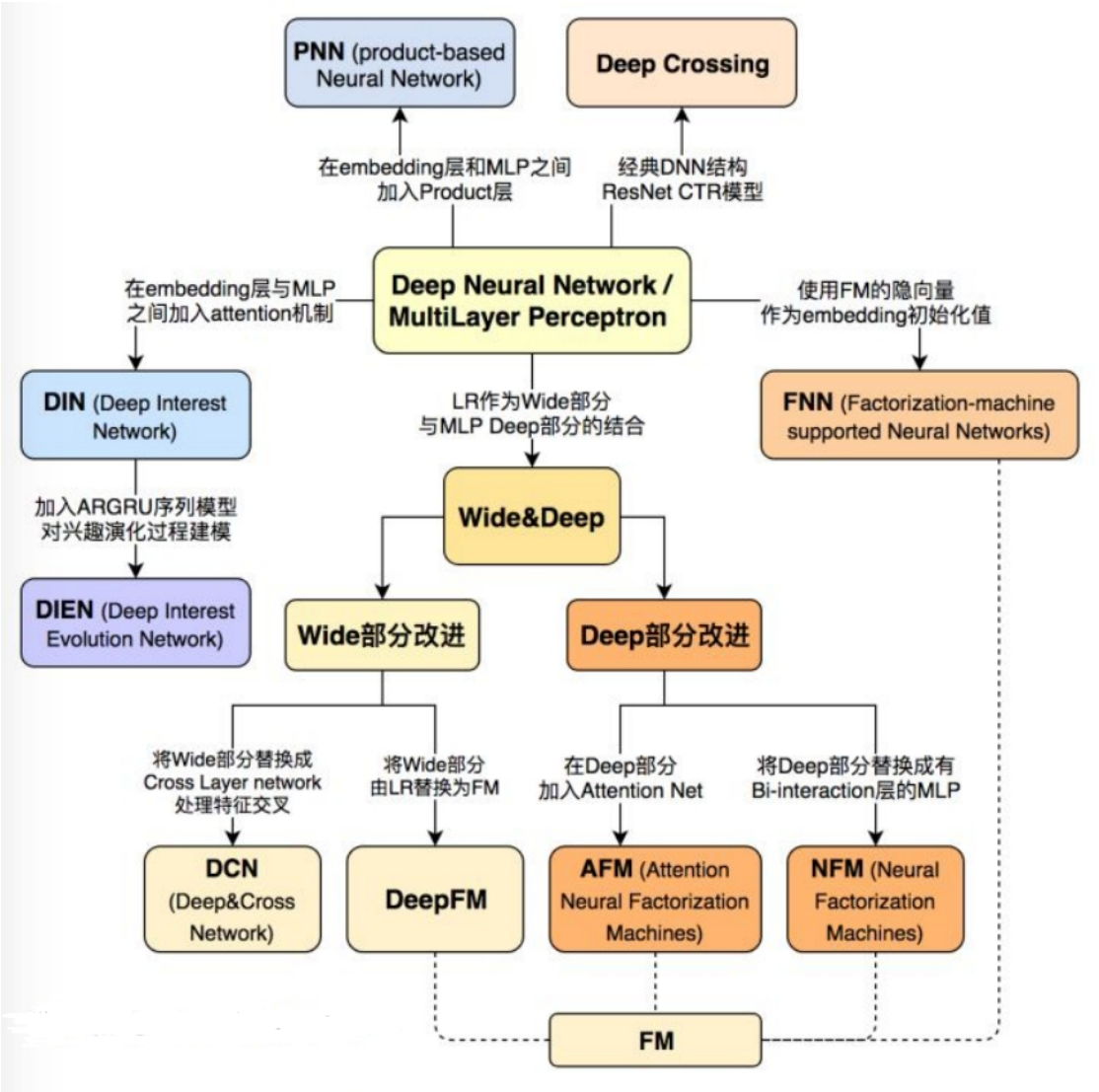
推荐算法与机器学习是密不可分的，推荐算法中有许多部分均可以引入机器学习里的算法．或许深度学习在推荐系统中没有像图像处理领域那般一枝独秀，但是依然能够在好几个方面起到不可替代的作用：直接从内容中提取特征，拥有较强表征能力；可以很方便地对噪声数据进行处理；可以更加准确地表示用户或物品的特征．

基于内容标签的推荐算法中自动化内容标签的生成，可以看作是机器学习领域的分类问题；推荐算法的召回阶段，可以建模成一个“超大规模多分类”问题；对内容的排序与评分，可以看作是一个回归问题；甚至在特征工程方面，深度学习也给我们提供了更为简单有效的方法，可以将我们从对特征工程复杂算法的研究中解脱出来．

我们以点击率（CTR）模型作为例子．CTR 预估模型在 2016 年被提出，计算广告和推荐系统领域全面进入了深度学习时代．现代的 CTR 模型，已经被 Google，微软，阿里等知名互联网公司成功应用．

目前基于机器学习与深度学习的推荐算法已经有了极大的发展．例如基于 DNN 的推荐算法，将推荐问题建模成一个“超大规模多分类问题”，通过数个隐层的 DNN 结构，实现大数据规模的召回阶段．再例如，基于 DeepFM 的推荐算法，其是一个集成了 FM 和 DNN 的神经网络框架．甚至，我们还能使用基于生成对抗网络（GAN）的推荐算法，同时训练生成器和判别器．

以下这张图片总结了目前深度学习在推荐系统领域常用的一系列模型．



3 应用案例

3.1 字节跳动

根据 36kr 的报道，2020 年 5 月字节跳动的估值已经超过 1000 亿美元，其中广告收入占比 85%，抖音贡献过半。至 2020 年 3 月，字节跳动已经有六万员工，并计划再增员一万人。投资人和内部消息将字节跳动 2019 年的营收定在 1,040 亿元至 1,400 亿元人民币，超过了 Uber、Snapchat 和推特的总和，广告收入也超越了腾讯、仅次于阿里巴巴。抖音的全球下载量达 1.15 亿次，固定用户近 10 亿。字节跳动的高速发展是惊人的。成立于 2012 年的字节跳动，在短短的八年间，便完成了从零到千亿美元市值的跨越，这出乎绝大部分人的意料。

我们可以看出，字节跳动的迅速发展的过程中很大一部分应该归功于字节跳动的技术，特别是字节跳动在推荐系统领域相关的技术与其多元的产品矩阵。

字节跳动公司的产品较为著名的有抖音，今日头条，西瓜视频等。抖音更是仅凭几个月的时间便风靡全国，甚至远洋海外，抖音的海外版本 TikTok 风靡全世界。我们能够在这三个产品中粗略看出字节跳动的发展方式：推荐系统 + 内容分发。

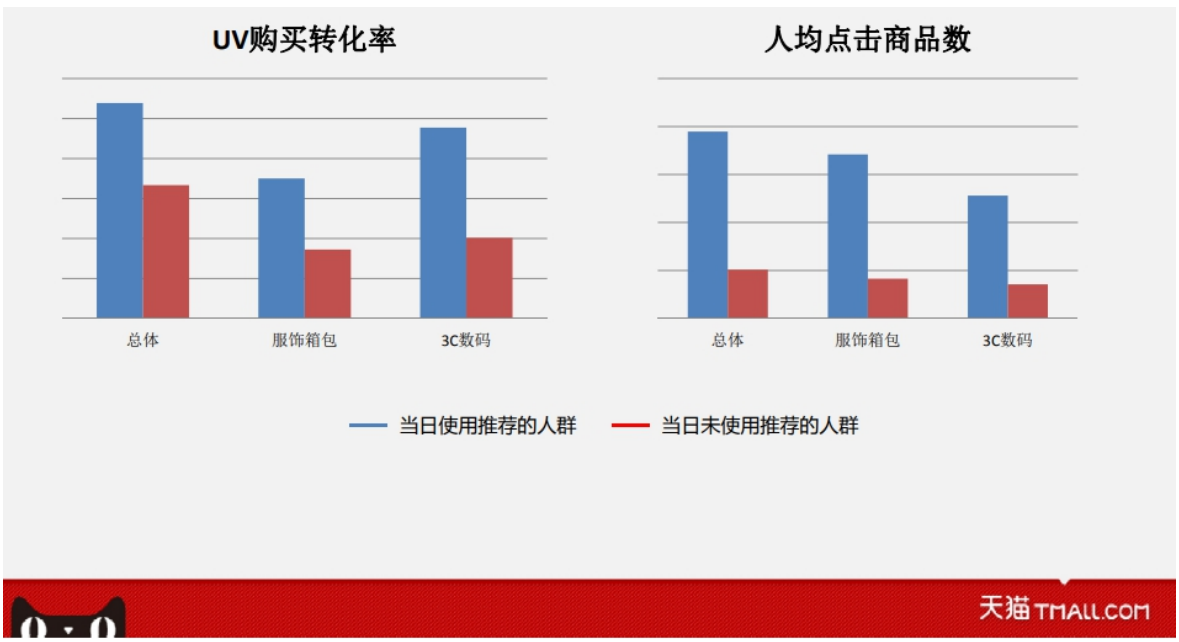
字节跳动成立之初，便依靠今日头条建立了“图文内容 + 内容分发 + 广告变现”的一条链商业模式。这条链条中的内容分发和广告变现均依靠推荐系统，用户画像等人工智能领域的内容。今日头条让创作者在其平台上创作，通过推荐系统分发给予创作者更高的流量，并且配套的广告变现系统让创作者的流量得以转化为报酬，从而激励了更多的创作者入驻今日头条平台。基于机器学习的推荐系统在里面发挥了不可磨灭的作用。

抖音在全球范围内发展火热，抖音代表的短视频 APP 的受众面远比人们想象中的广。类似于今日头条，抖音也是采用了“短视频内容 + 内容分发 + 广告变现”这一类似的模式，其中的核心仍然是一个优质的推荐系统分发内容，进而加入广告变现。

3.2 电商平台

如今的电商平台，例如亚马逊和淘宝，均拥有“猜你喜欢”模块，基于推荐系统和用户过往的浏览数据购物数据进行个性化的商品推荐.

阿里巴巴的天猫平台通过推荐系统也成功提高了相当程度的购物比例.



4 个人思考

推荐系统是当下互联网时代不可或缺的一部分. 由于当前互联网行业的迅猛发展, 与用户个性化的需求, 商品需求的长尾效应等等因素影响下, 各大平台均要向着推荐系统的方向转型. 基于推荐系统的新型新闻平台取代了中心化的新闻平台, 基于推荐系统的电商平台获得了更高的购物比率, 基于推荐系统的 UGC 平台给了每个人表达自我的机会. 这也说明了, 在新的互联网时代中, “个性化” 的力量.

但是每一样事物往往都是双刃剑, 推荐系统的发展也带来了一定的弊端.

推荐系统首先涉及到的就是隐私问题. 想要拥有一个优秀的推荐系统, 对用户数据的收集必不可少. 但是各大互联网平台是否拥有对用户数据进行收集的权利? 例如用户所在位置隐私, 人们总是不希望自己当前所在位置被泄漏, 但是许多的 APP 却希望能够获取用户当前所在位置, 以更精确地进行广告推送. 于是边产生了用户与平台之间的隐私矛盾.

推荐系统另一个设计到的问题是信息茧房. 由于推荐系统的存在, 每个人会被迫性地更加地倾向于接收同一类的信息, 也即推荐系统认为用户会感兴趣的信息. 于是用户就在这个过程中产生了信息偏差, 认为自己所接收的信息代表了这个社会的主流信息, 渐渐在思想上作茧自缚, 不迈出自己所在的小圈子半步. 由于这种信息茧房的影响, 社会会趋向于分成不同的小群体小圈子, 而不同小圈子的对立会日益严重, 严重到一定程度, 甚至会导致不同圈子的对立与争吵, 例如阶级对立, 性别对立等等, 影响到社会稳定.

因此, 我们身为人工智能专业的研究者, 要考虑到这不同的方面, 权衡技术发展与社会影响. 在努力发展基于机器学习的推荐系统基础上, 也要加强自己的社会责任心, 造就一个更好的社会.

References:

[1] Deep Learning 可以用来做推荐系统吗 https://www.zhihu.com/question/20830906/answer/681688041
[2] 推荐系统 - 维基百科 https://zh.wikipedia.org/wiki/%E6%8E%A8%E8%96%A6%E7%B3%BB%E7%B5%B1
[3] 字节跳动的发展潜力 https://www.zhihu.com/question/355576724/answer/1303375369
[4] 特征工程到底是什么 https://www.zhihu.com/question/29316149/answer/110159647
[5] 什么是推荐系统 https://zhuanlan.zhihu.com/p/27126285
[6] 《推荐系统与深度学习》 黄昕等著
[7] 字节跳动 - 维基百科 https://zh.wikipedia.org/wiki/%E5%AD%97%E8%8A%82%E8%B7%B3%E5%8A%A8
[8] 天猫推荐业务与算法架构 https://topic.it168.com/factory/adc2013/doc/zhangqi.pdf
[9] 国内外在推荐系统领域的发展现状？ https://www.zhihu.com/question/29531839