# **跳动的字节**

### 推荐系统中的人工智能

### 一．迅速发展的字节跳动

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。

所以字节跳动到底是怎么达成这个成就的？我们可以很容易地看出，这其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，还有字节跳动多元的产品矩阵。

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：AI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AI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。AI在里面发挥了不可磨灭的作用。

2016年底，字节跳动下属的短视频平台，抖音，开始在用户市场发力。在BAT没有太过注意的赛道上，字节跳动完成了弯道超车。即使腾讯想尽一切办法，推出了微视，视频号等等一系列APP，均难以挽回败局。抖音在全球范围内变得火热，短视频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。

那么处于字节跳动技术核心的推荐系统，究竟与人工智能有着那些联系呢？

### 二．用于内容分发的推荐系统

#### 1. 什么是推荐系统

什么是推荐系统？维基百科将推荐系统定义为一种信息过滤系统，用于预测用户对物品的“评分”或“偏好”。这里的物品是非常多元化的，对于电商行业来说，这个物品可以是你要售卖的商品；对于社交网站来说，这个物品可以是用户感兴趣的人或群组；对于内容分发来说，这个物品可以是如图文音乐视频等多媒体内容；对于广告行业来说，这个物品还可以是众多的广告。就以字节跳动的推荐系统来说，其主要应用场景便是多媒体内容和广告。

#### 常见的推荐算法

推荐算法有很多种，常见的就有基于内容标签的推荐算法和协同过滤算法两种。

基于内容标签的算法是最为传统的，你可以在各大内容网站看到它们的身影，例如商品的类别便签，新闻的分类标签，音乐的流派便签等，然后结合用户的历史行为，便能进行简单而有效的推荐。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。在现代，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。

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。

#### 技术具体内容

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。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，更要看身高数据。但是如果给你一个人具体的身高和体重，你能很直接地判断出这个人胖不胖吗？并不能。但是我们通过引入BMI指数，BMI=体重/(身高^2)。我们通过BMI指数便能非常清晰地看出这个人身材如何。

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。为什么要进行排序，像召回步骤那样直接去除低相关性的内容不行吗？事实上的确不行，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。

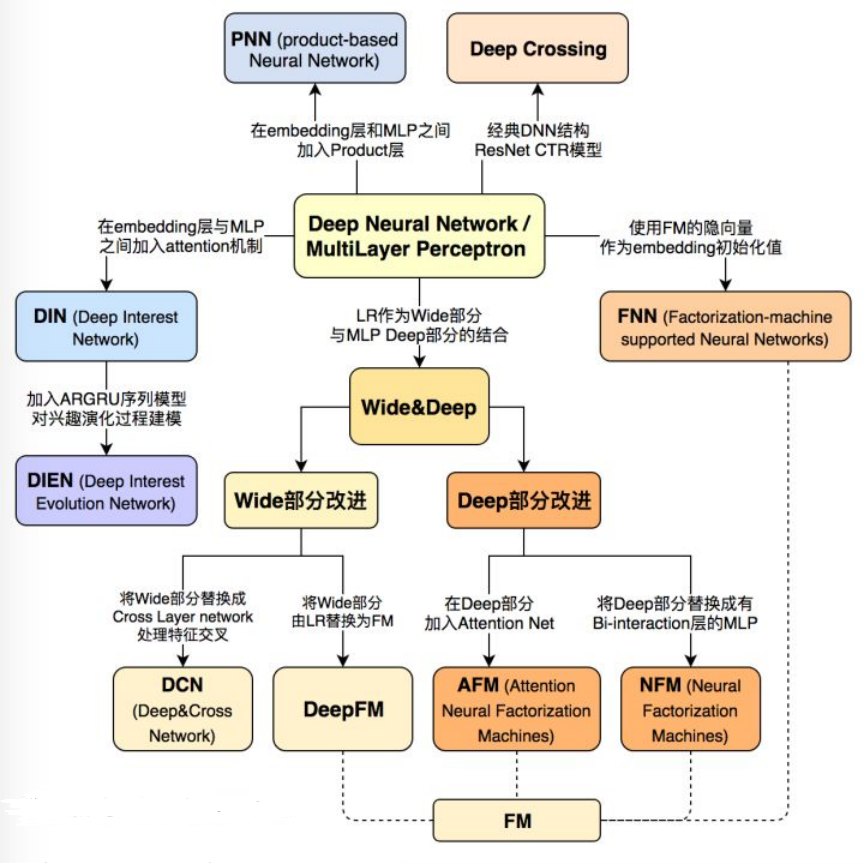
还有一些其他的相关技术，例如冷启动问题。一个新用户，新内容加入推荐系统时，系统只有很少于其相关的信息，甚至没有与其相关的信息，这被称为冷启动问题。我们要考虑相关的解决方案，例如推荐被绝大多数人喜欢的内容，在用户注册时让其选择相关的标签等方式。

### 人工智能与推荐系统

人工智能和推荐系统是密不可分的。推荐系统实现的过程，其实质就是一个对人工智能在推荐领域的应用过程。其中涉及到的矩阵论，概率论等数学方法都是人工智能的基石。而近来人工智能领域深度学习的发展也带来了推荐系统的一次革新。

自动化内容便签的生成，可以看作是机器学习领域的分类问题；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，深度学习也提供了一个更为简单有效的方法，将人们从对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，是工程导向的，而不是仅用实验数据验证或学术创新用的。



总而言之，人工智能在内容分发领域已经被广泛地应用。而字节跳动正是人工智能驱动内容分发领域地佼佼者，也凭借相关的技术取得了行业领先的地位。而目前推荐系统对人工智能的应用还远远未到尽头，人工智能领域还有着更为广阔的未来。

**参考文献：**

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**