今天，算法分发已经是信息平台、搜索引擎、浏览器、社交软件等几乎所有软件的标配，但同时，算法也开始面临质疑、挑战和误解。今日头条的推荐算法，从 2012 年 9 月第一版开发运行至今，已经经过四次大的调整和修改。

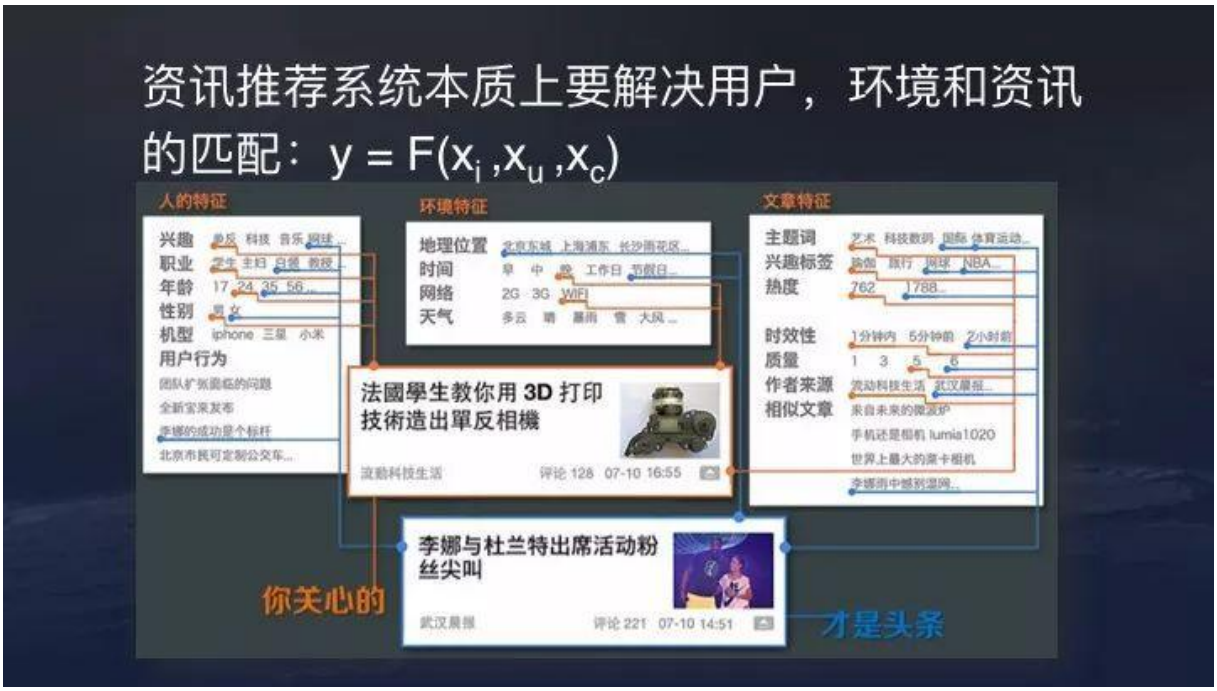
今日头条委托资深算法架构师曹欢欢博士，公开今日头条的算法原理，以期推动整个行业问诊算法、建言算法；通过让算法透明，来消除各界对算法的误解，并逐步推动整个行业让算法更好的造福社会。

以下为《今日头条算法原理》全文。

今日头条资深算法架构师曹欢欢

本次分享将主要介绍今日头条推荐系统概览以及内容分析、用户标签、评估分析，内容安全等原理。

一、系统概览

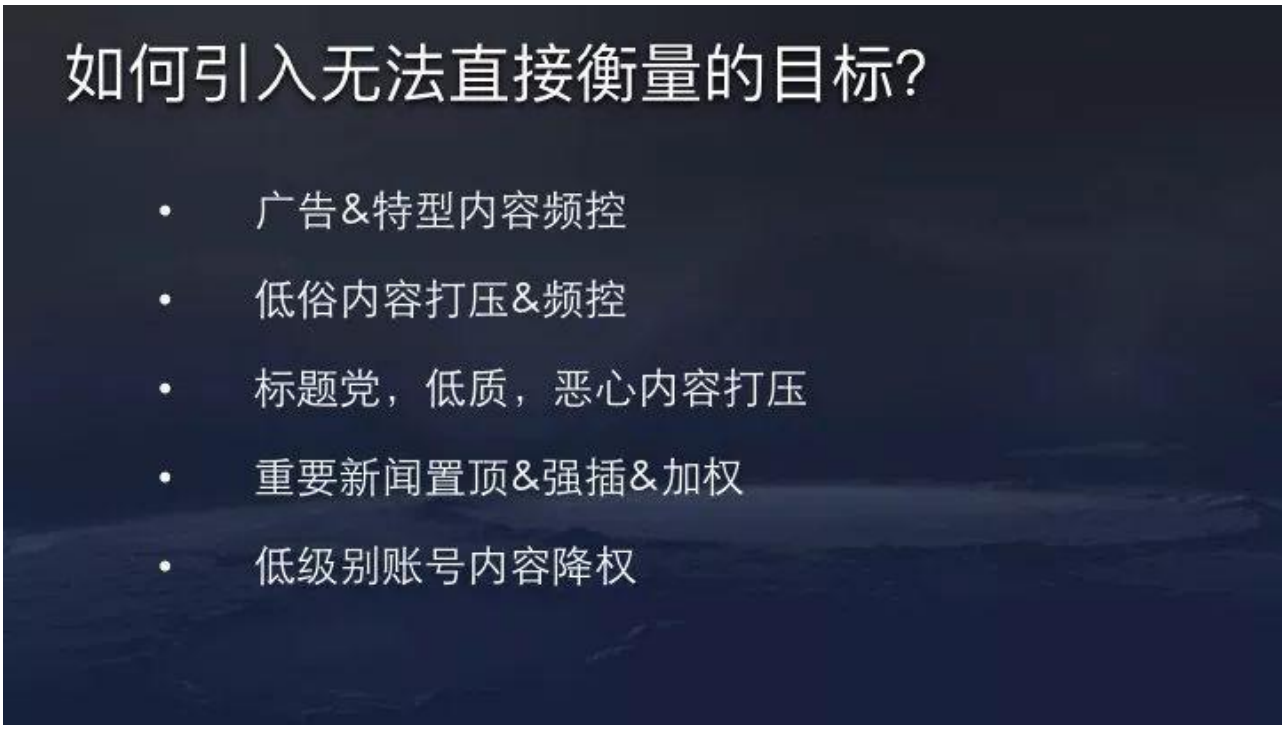


推荐系统，如果用形式化的方式去描述实际上是拟合一个用户对内容满意度的函数，这个函数需要输入三个维度的变量。第一个维度是内容。头条现在已经是一个综合内容平台，图文、视频、UGC 小视频、问答、微头条，每种内容有很多自己的特征，需要考虑怎样提取不同内容类型的特征做好推荐。第二个维度是用户特征。包括各种兴趣标签，职业、年龄、性别等，还有很多模型刻划出的隐式用户兴趣等。第三个维度是环境特征。这是移动互联网时代推荐的特点，用户随时随地移动，

在工作场合、通勤、旅游等不同的场景，信息偏好有所偏移。结合三方面的维度，模型会给出一个预估，即推测推荐内容在这一场景下对这一用户是否合适。

这里还有一个问题，如何引入无法直接衡量的目标？

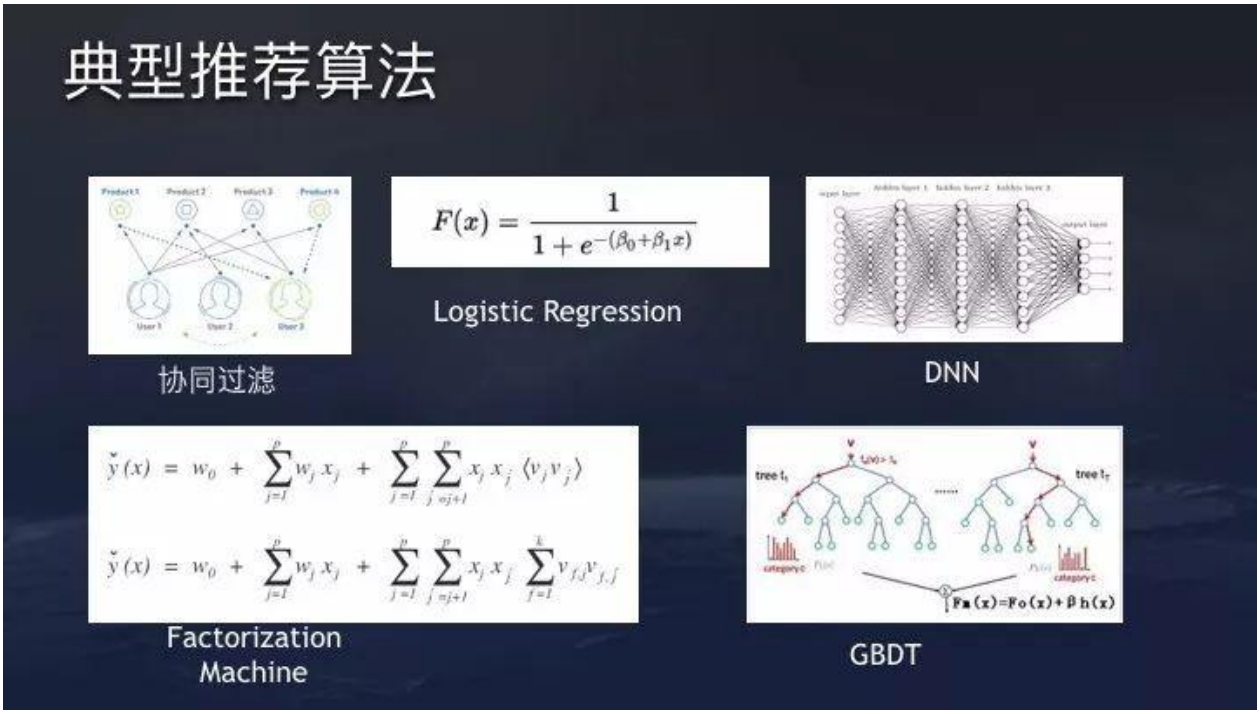
推荐模型中，点击率、阅读时间、点赞、评论、转发包括点赞都是可以量化的目标，能够用模型直接拟合做预估，看线上提升情况可以知道做的好不好。但一个大体量的推荐系统，服务用户众多，不能完全由指标评估，引入数据指标以外的要素也很重要。



比如广告和特型内容频控。像问答卡片就是比较特殊的内容形式，其推荐的目标不完全是让用户浏览，还要考虑吸引用户回答为社区贡献内容。这些内容和普通内容如何混排，怎样控制频控都需要考虑。

此外，平台出于内容生态和社会责任的考量，像低俗内容的打压，标题党、低质内容的打压，重要新闻的置顶、加权、强插，低级别账号内容降权都是算法本身无法完成，需要进一步对内容进行干预。

下面我将简单介绍在上述算法目标的基础上如何对其实现。



前面提到的公式 y = F(Xi ,Xu ,Xc)，是一个很经典的监督学习问题。可实现的方法有很多，比如传统的协同过滤模型，监督学习算法 Logistic Regression 模型，基于深度学习的模型，Factorization Machine 和 GBDT 等。

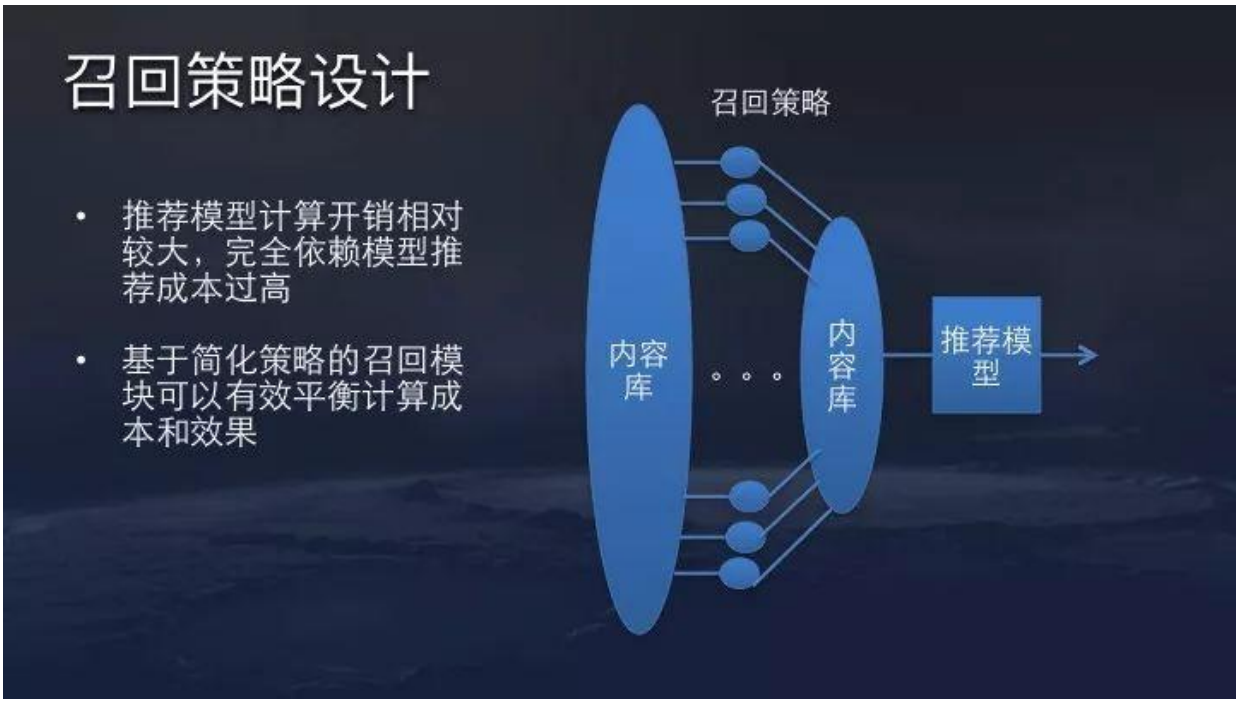
一个优秀的工业级推荐系统需要非常灵活的算法实验平台，可以支持多种算法组合，包括模型结构调整。因为很难有一套通用的模型架构适用于所有的推荐场景。现在很流行将 LR 和 DNN 结合，前几年 Facebook 也将 LR 和 GBDT 算法做结合。今日头条旗下几款产品都在沿用同一套强大的算法推荐系统，但根据业务场景不同，模型架构会有所调整。

模型之后再看一下典型的推荐特征，主要有四类特征会对推荐起到比较重要的作用。

- 第一类是相关性特征，就是评估内容的属性和与用户是否匹配。显性的匹配包括关键词匹配、分类匹配、来源匹配、主题匹配等。像 FM 模型中也有一些隐性匹配，从用户向量与内容向量的距离可以得出。
- 第二类是环境特征，包括地理位置、时间。这些既是 bias 特征，也能以此构建一些匹配特征。
- 第三类是热度特征。包括全局热度、分类热度，主题热度，以及关键词热度等。内容热度信息在大的推荐系统特别在用户冷启动的时候非常有效。
- 第四类是协同特征，它可以在部分程度上帮助解决所谓算法越推越窄的问题。协同特征并非考虑用户已有历史。而是通过用户行为分析不同用户间相似性，比如点击相似、兴趣分类相似、主题相似、兴趣词相似，甚至向量相似，从而扩展模型的探索能力。

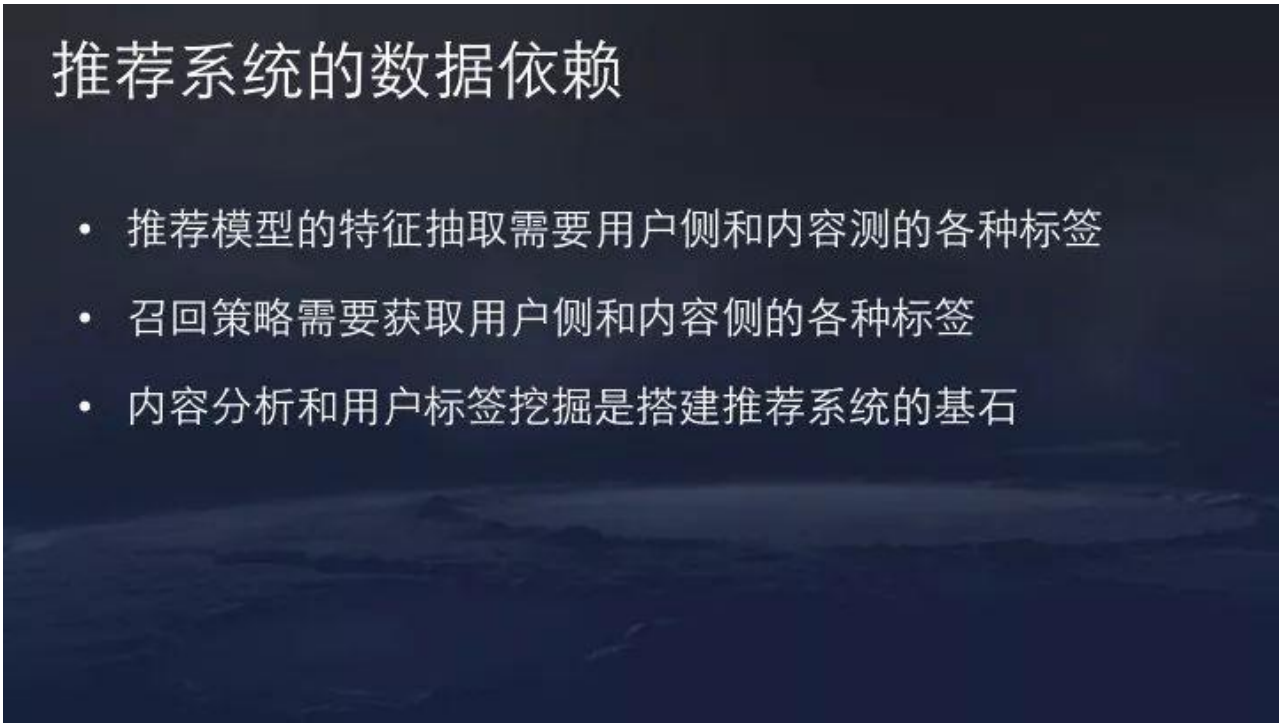
模型的训练上，头条系大部分推荐产品采用实时训练。实时训练省资源并且反馈快，这对信息流产品非常重要。用户需要行为信息可以被模型快速捕捉并反馈至下一刷的推荐效果。我们线上目前基于 storm 集群实时处理样本数据，包括点击、展现、收藏、分享等动作类型。模型参数服务器是内部开发的一套高性能的系统，因为头条数据规模增长太快，类似的开源系统稳定性和性能无法满足，而我们自研的系统底层做了很多针对性的优化，提供了完善运维工具，更适配现有的业务场景。

目前，头条的推荐算法模型在世界范围内也是比较大的，包含几百亿原始特征和数十亿向量特征。整体的训练过程是线上服务器记录实时特征，导入到 Kafka 文件队列中，然后进一步导入 Storm 集群消费 Kafka 数据，客户端回传推荐的 label 构造训练样本，随后根据最新样本进行在线训练更新模型参数，最终线上模型得到更新。这个过程中主要的延迟在用户的动作反馈延时，因为文章推荐后用户不一定马上看，不考虑这部分时间，整个系统是几乎实时的。



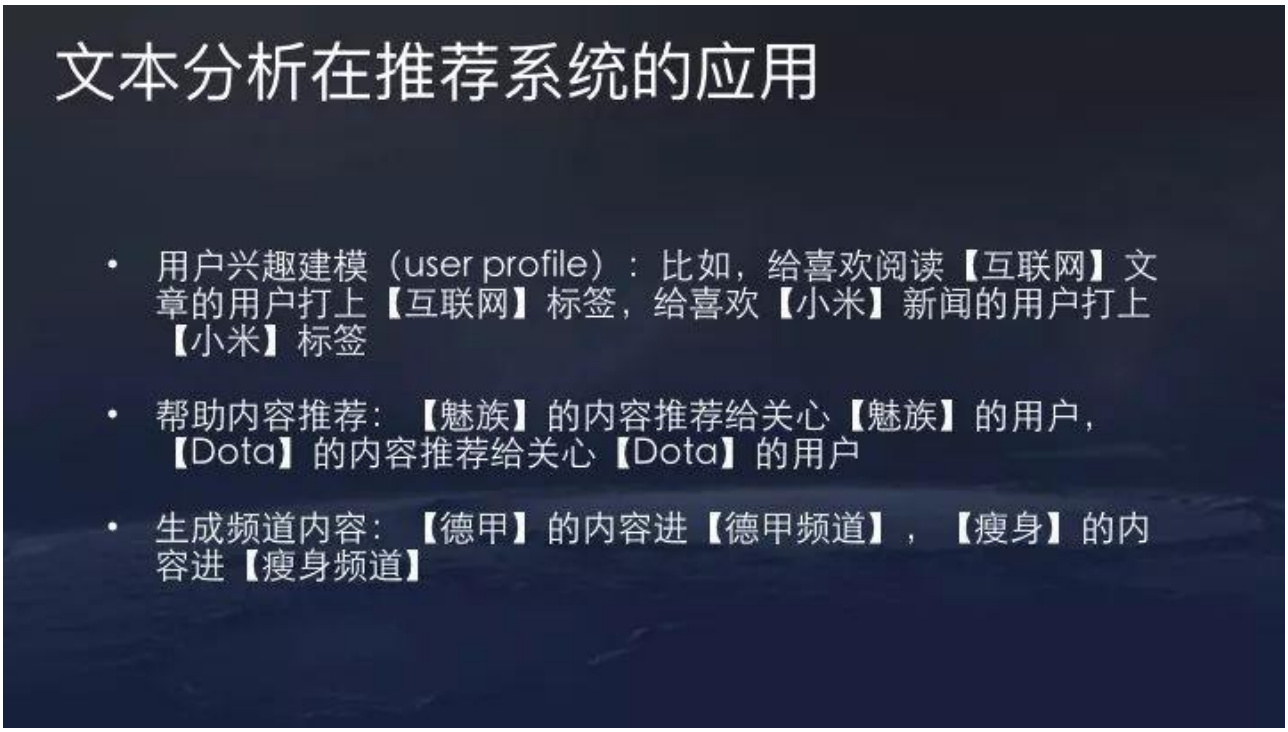
但因为头条目前的内容量非常大，加上小视频内容有千万级别，推荐系统不可能所有内容全部由模型预估。所以需要设计一些召回策略，每次推荐时从海量内容中筛选出千级别的内容库。召回策略最重要的要求是性能要极致，一般超时不能超过 50 毫秒。

召回策略种类有很多，我们主要用的是倒排的思路。离线维护一个倒排，这个倒排的 key 可以是分类，topic，实体，来源等，排序考虑热度、新鲜度、动作等。线上召回可以迅速从倒排中根据用户兴趣标签对内容做截断，高效的从很大的内容库中筛选比较靠谱的一小部分内容。



二、内容分析

内容分析包括文本分析，图片分析和视频分析。头条一开始主要做资讯，今天我们主要讲一下文本分析。文本分析在推荐系统中一个很重要的作用是用户兴趣建模。没有内容及文本标签，无法得到用户兴趣标签。举个例子，只有知道文章标签是互联网，用户看了互联网标签的文章，才能知道用户有互联网标签，其他关键词也一样。

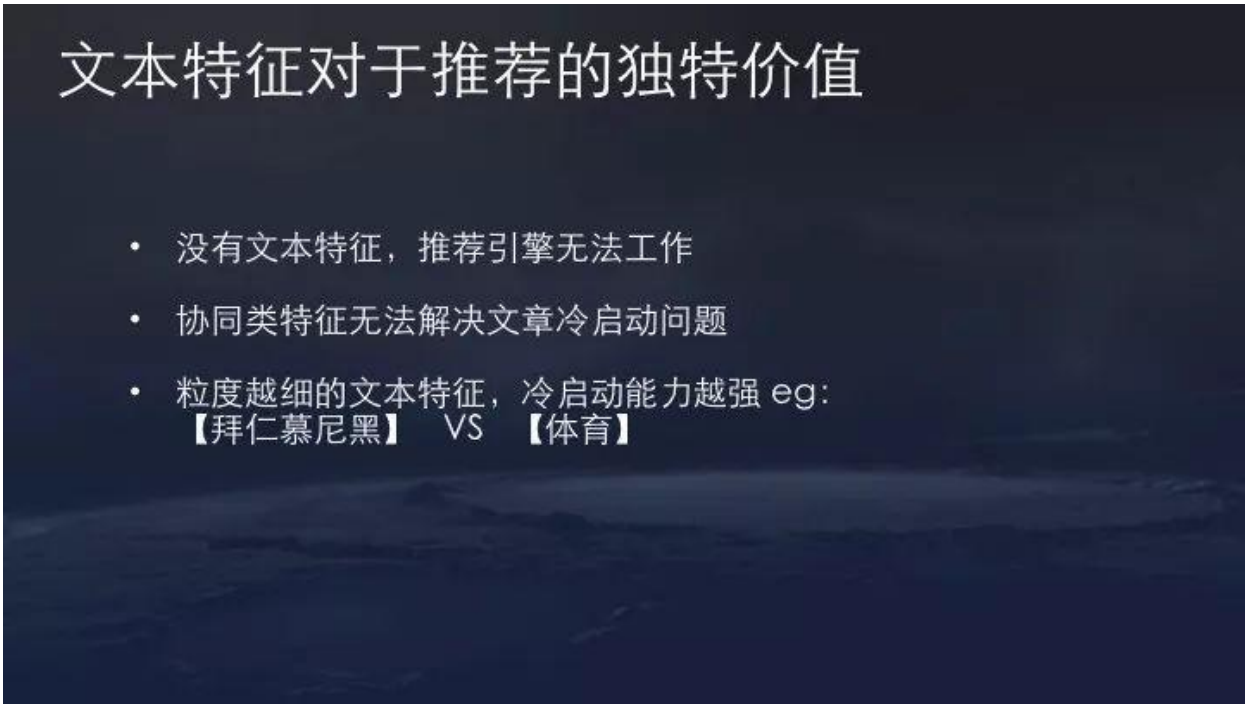


另一方面，文本内容的标签可以直接帮助推荐特征，比如魅族的内容可以推荐给关注魅族的用户，这是用户标签的匹配。如果某段时间推荐主频道效果不理想，出现推荐窄化，用户会发现到具体的频道推荐（如科技、体育、娱乐、军事等）中阅读后，再回主 feed,推荐效果会更好。因为整个模型是打通的，子频道探索空间较小，更容易满足用户需求。只通过单一信道反馈提高推荐准确率难度会比较大，子频道做的好很重要。而这也需要好的内容分析。

上图是今日头条的一个实际文本 case。

可以看到，这篇文章有分类、关键词、topic、实体词等文本特征。当然不是没有文本特征，推荐系统就不能工作，推荐系统最早期应用在 Amazon,甚至沃尔玛时代就有，包括 Netfilx 做视频推荐也没有文本特征直接协同过滤推荐。但对资讯类产品而言，大部分是消费当天内容，没有文本特征新内容冷启动非常困难，协同类特征无法解决文章冷启动问题。

今日头条推荐系统主要抽取的文本特征包括以下几类。首先是语义标签类特征，显式为文章打上语义标签。这部分标签是由人定义的特征，每个标签有明确的意义，标签体系是预定义的。此外还有隐式语义特征，主要是 topic 特征和关键词特征，其中 topic 特征是对于词概率分布的描述，无明确意义；而关键词特征会基于一些统一特征描述，无明确集合。

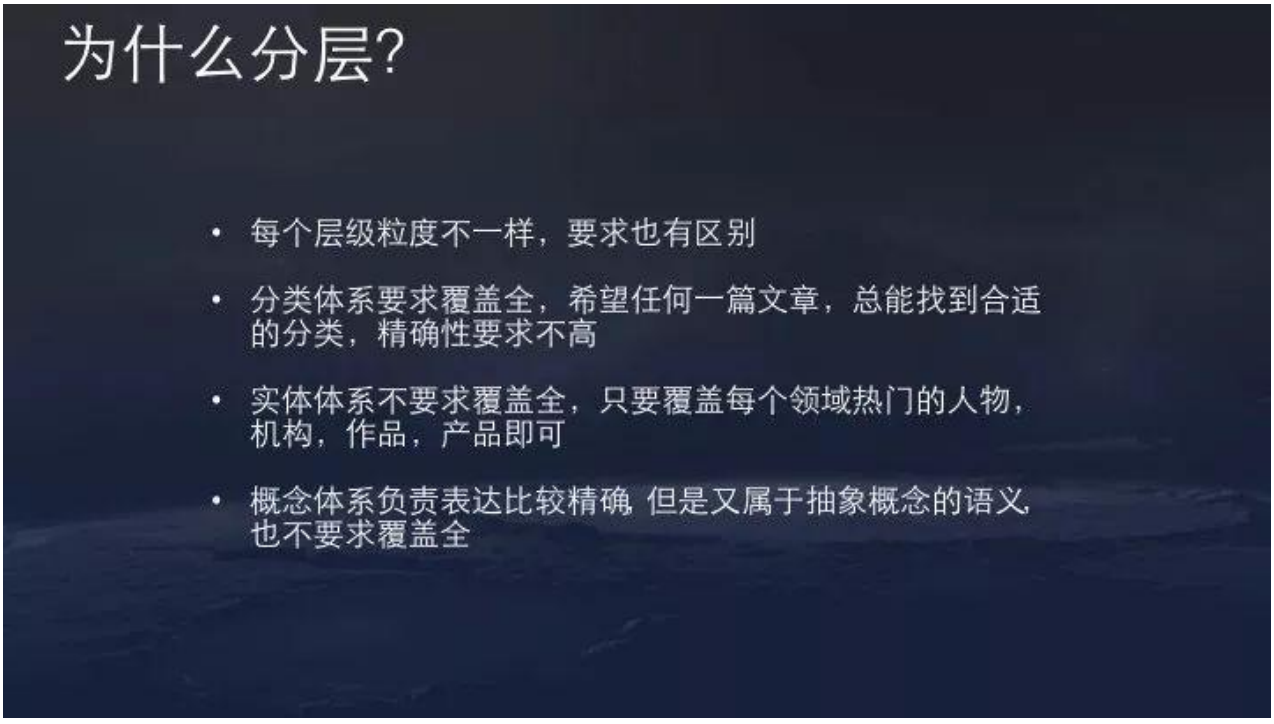


另外文本相似度特征也非常重要。在头条，曾经用户反馈最大的问题之一就是为什么总推荐重复的内容。这个问题的难点在于，每个人对重复的定义不一样。举个例子，有人觉得这篇讲皇马和巴萨的文章，昨天已经看过类似内容，今天还说这两个队那就是重复。但对于一个重度球迷而言，尤其

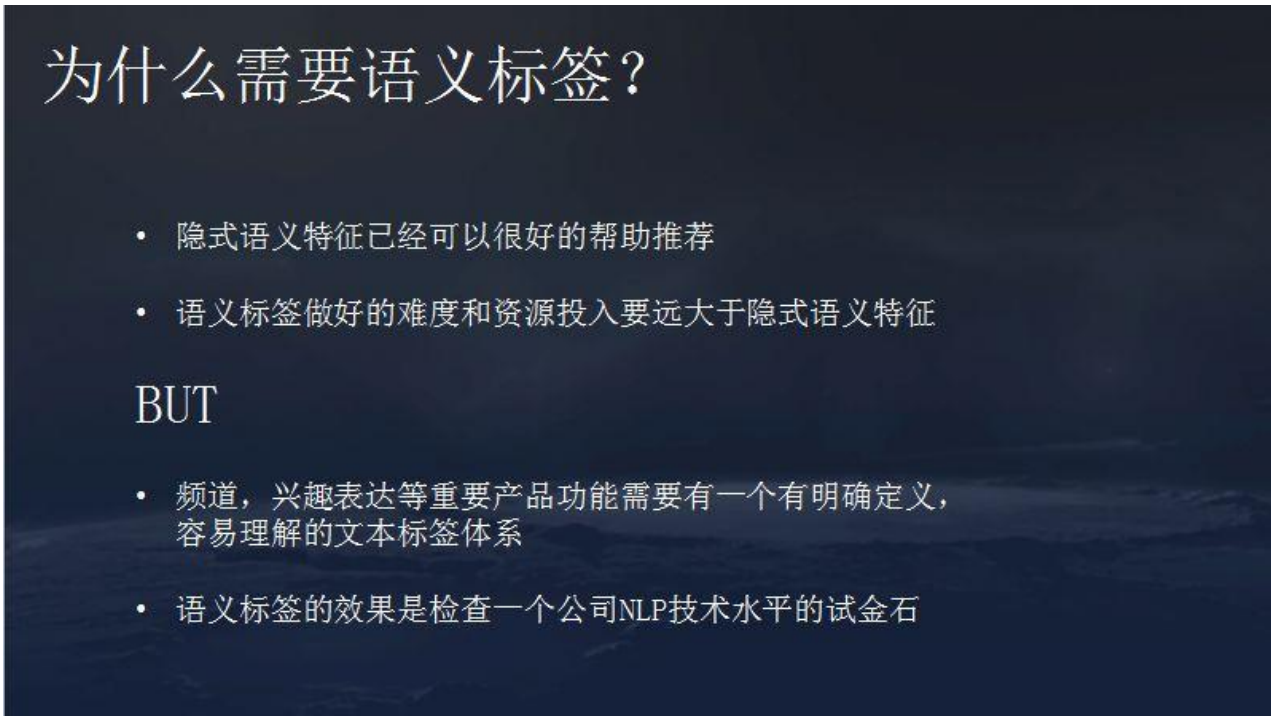
是巴萨的球迷，恨不得所有报道都看一遍。解决这一问题需要根据判断相似文章的主题、行文、主体等内容，根据这些特征做线上策略。

同样，还有时空特征，分析内容的发生地点以及时效性。比如武汉限行的事情推给北京用户可能就没有意义。最后还要考虑质量相关特征，判断内容是否低俗，色情，是否是软文，鸡汤？

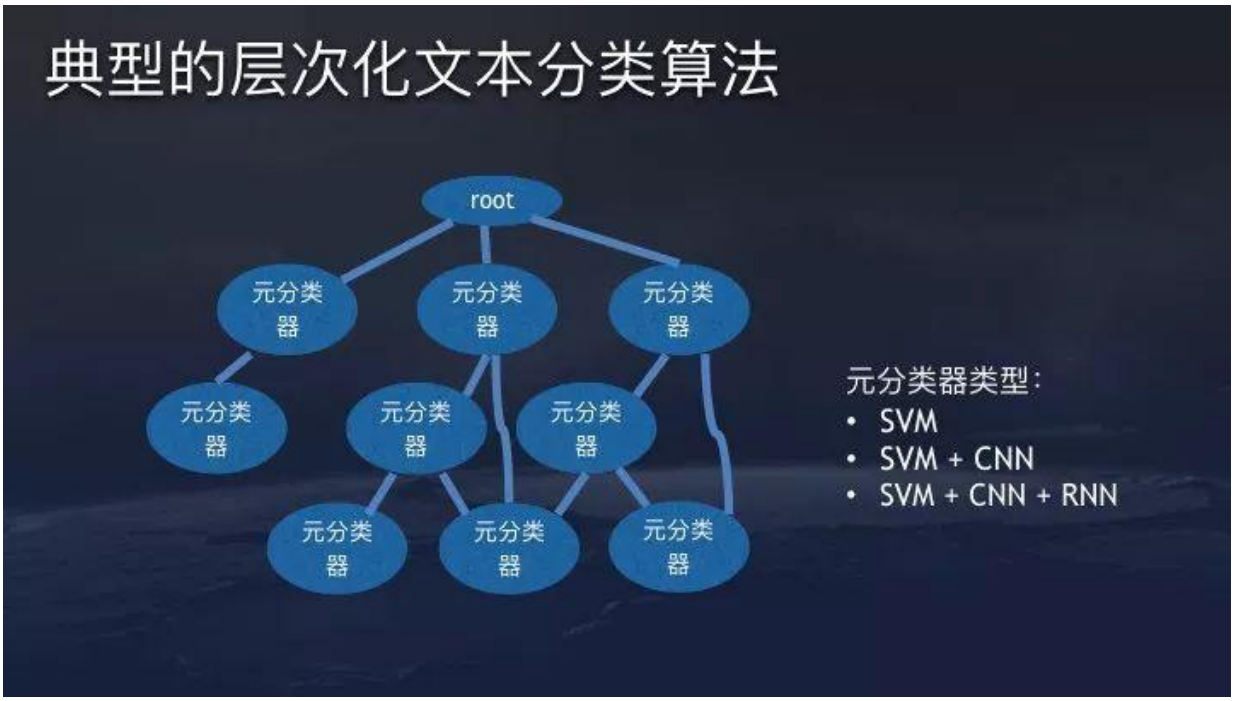
上图是头条语义标签的特征和使用场景。他们之间层级不同，要求不同。



分类的目标是覆盖全面，希望每篇内容每段视频都有分类；而实体体系要求精准，相同名字或内容要能明确区分究竟指代哪一个人或物，但不用覆盖很全。概念体系则负责解决比较精确又属于抽象概念的语义。这是我们最初的分类，实践中发现分类和概念在技术上能互用，后来统一用了一套技术架构。



目前，隐式语义特征已经可以很好的帮助推荐，而语义标签需要持续标注，新名词新概念不断出现，标注也要不断迭代。其做好的难度和资源投入要远大于隐式语义特征，那为什么还需要语义标签？有一些产品上的需要，比如频道需要有明确定义的分类内容和容易理解的文本标签体系。语义标签的效果是检查一个公司 NLP 技术水平的试金石。

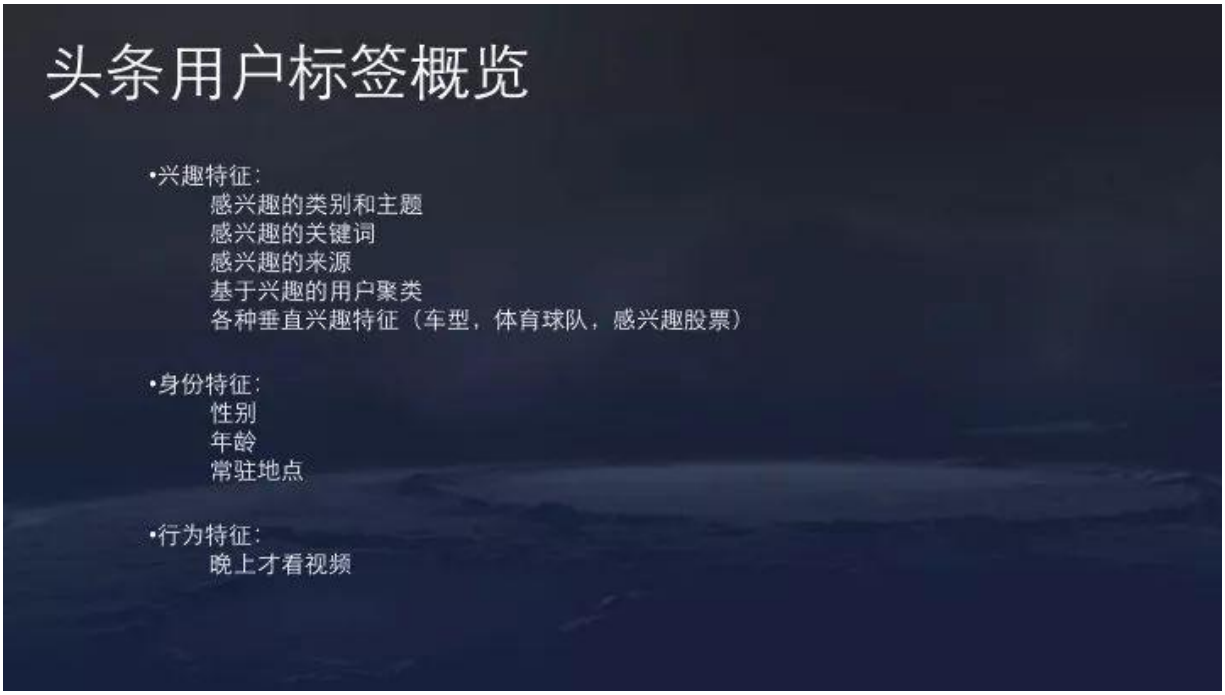


今日头条推荐系统的线上分类采用典型的层次化文本分类算法。最上面 Root，下面第一层的分类是像科技、体育、财经、娱乐，体育这样的大类，再下面细分足球、篮球、乒乓球、网球、田径、游泳等，足球再细分国际足球、中国足球，中国足球又细分中甲、中超、国家队等，相比单独的分类器，利用层次化文本分类算法能更好地解决数据倾斜的问题。有一些例外是，如果要提高召回，可以看到我们连接了一些飞线。这套架构通用，但根据不同的问题难度，每个元分类器可以异构，像有些分类 SVM 效果很好，有些要结合 CNN，有些要结合 RNN 再处理一下。

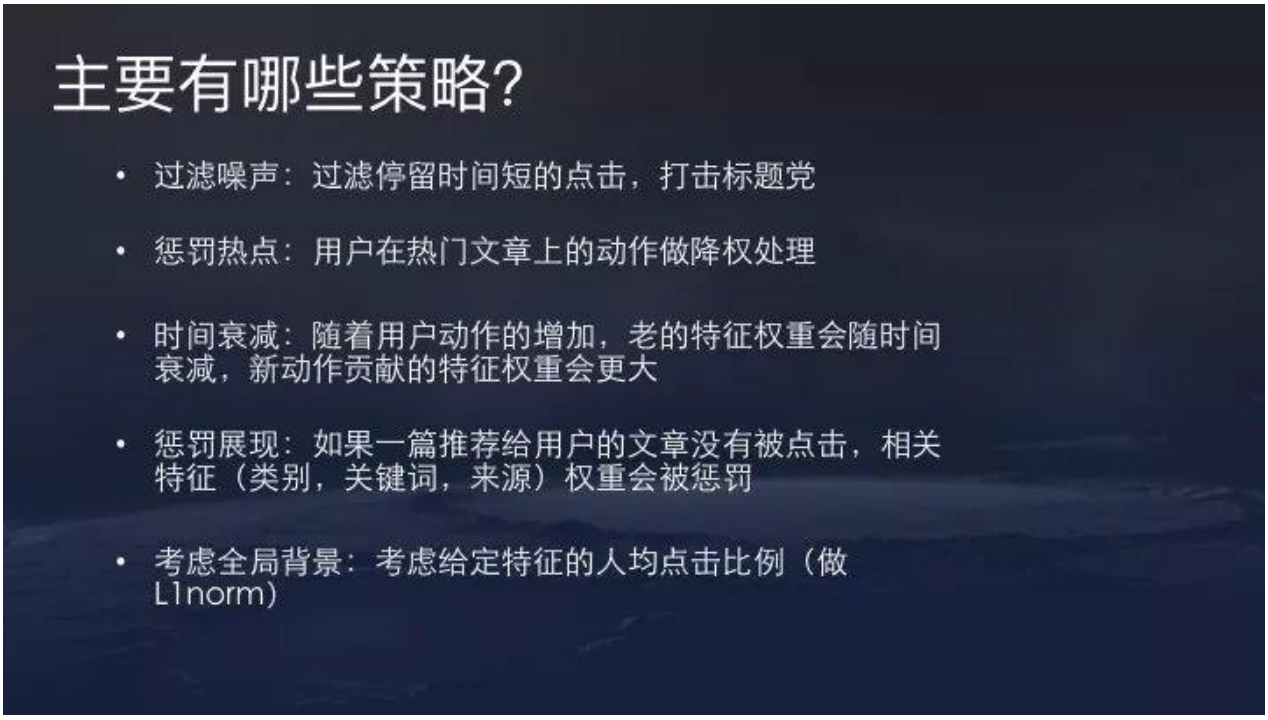
上图是一个实体词识别算法的 case。基于分词结果和词性标注选取候选，期间可能需要根据知识库做一些拼接，有些实体是几个词的组合，要确定哪几个词结合在一起能映射实体的描述。如果结果映射多个实体还要通过词向量、topic 分布甚至词频本身等去歧，最后计算一个相关性模型。

三、用户标签

内容分析和用户标签是推荐系统的两大基石。内容分析涉及到机器学习的内容多一些，相比而言，用户标签工程挑战更大。



今日头条常用的用户标签包括用户感兴趣的类别和主题、关键词、来源、基于兴趣的用户聚类以及各种垂直兴趣特征（车型，体育球队，股票等）。还有性别、年龄、地点等信息。性别信息通过用户第三方社交账号登录得到。年龄信息通常由模型预测，通过机型、阅读时间分布等预估。常驻地点来自用户授权访问位置信息，在位置信息的基础上通过传统聚类的方法拿到常驻点。常驻点结合其他信息，可以推测用户的工作地点、出差地点、旅游地点。这些用户标签非常有助于推荐。

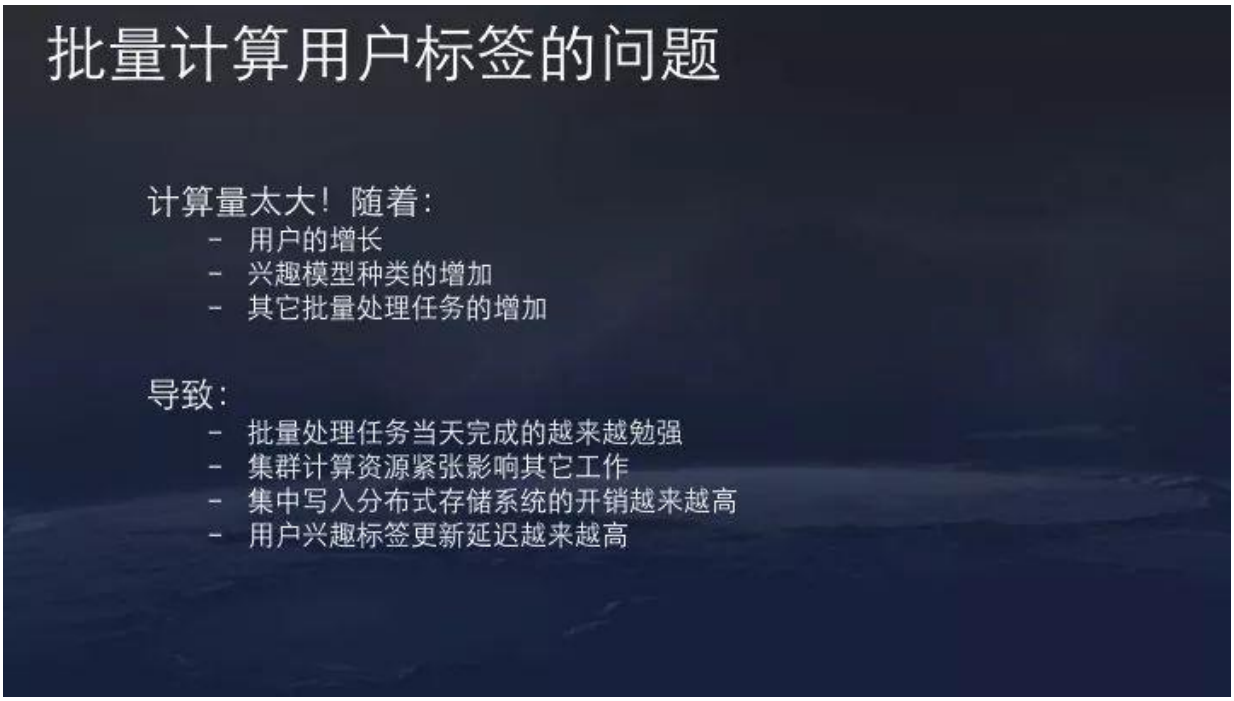


当然最简单的用户标签是浏览过的内容标签。但这里涉及到一些数据处理策略。主要包括：

- 过滤噪声。通过停留时间短的点击，过滤标题党。
- 热点惩罚。对用户在一些热门文章（如前段时间 PG One 的新闻）上的动作做降权处理。理论上，传播范围较大的内容，置信度会下降。
- 时间衰减。用户兴趣会发生偏移，因此策略更偏向新的用户行为。因此，随着用户动作的增加，老的特征权重会随时间衰减，新动作贡献的特征权重会更大。
- 惩罚展现。如果一篇推荐给用户的文章没有被点击，相关特征（类别，关键词，来源）权重会被惩罚。

当然同时，也要考虑全局背景，是不是相关内容推送比较多，以及相关的关闭和 dislike 信号等。

用户标签挖掘总体比较简单，主要还是刚刚提到的工程挑战。头条用户标签第一版是批量计算框架，流程比较简单，每天抽取昨天的日活用户过去两个月的动作数据，在 Hadoop 集群上批量计算结果。



但问题在于，随着用户高速增长，兴趣模型种类和其他批量处理任务都在增加，涉及到的计算量太大。2014 年，批量处理任务几百万用户标签更新的 Hadoop 任务，当天完成已经开始勉强。集群计算资源紧张很容易影响其它工作，集中写入分布式存储系统的压力也开始增大，并且用户兴趣标签更新延迟越来越高。

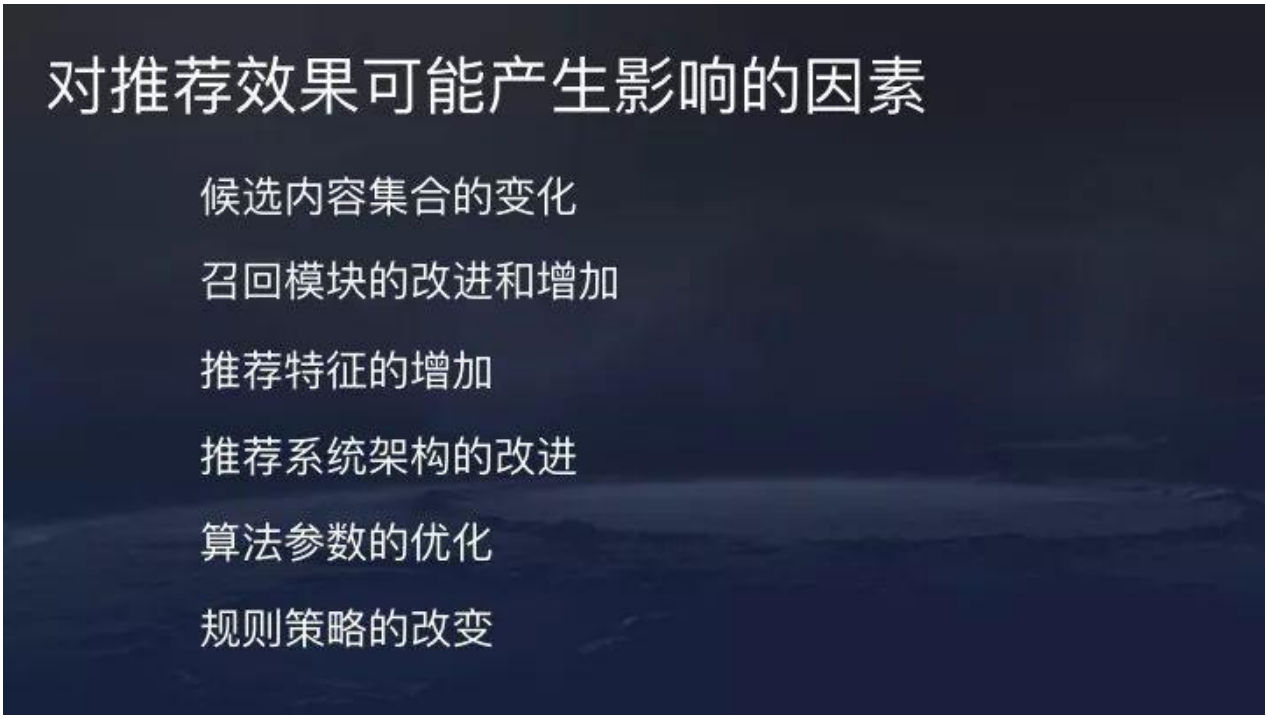
面对这些挑战。2014 年底今日头条上线了用户标签 Storm 集群流式计算系统。改成流式之后，只要有用户动作更新就更新标签，CPU 代价比较小，可以节省 80%的 CPU 时间，大大降低了计算资源开销。同时，只需几十台机器就可以支撑每天数千万用户的兴趣模型更新，并且特征更新速度非常快，基本可以做到准实时。这套系统从上线一直使用至今。

当然，我们也发现并非所有用户标签都需要流式系统。像用户的性别、年龄、常驻地点这些信息，不需要实时重复计算，就仍然保留 daily 更新。

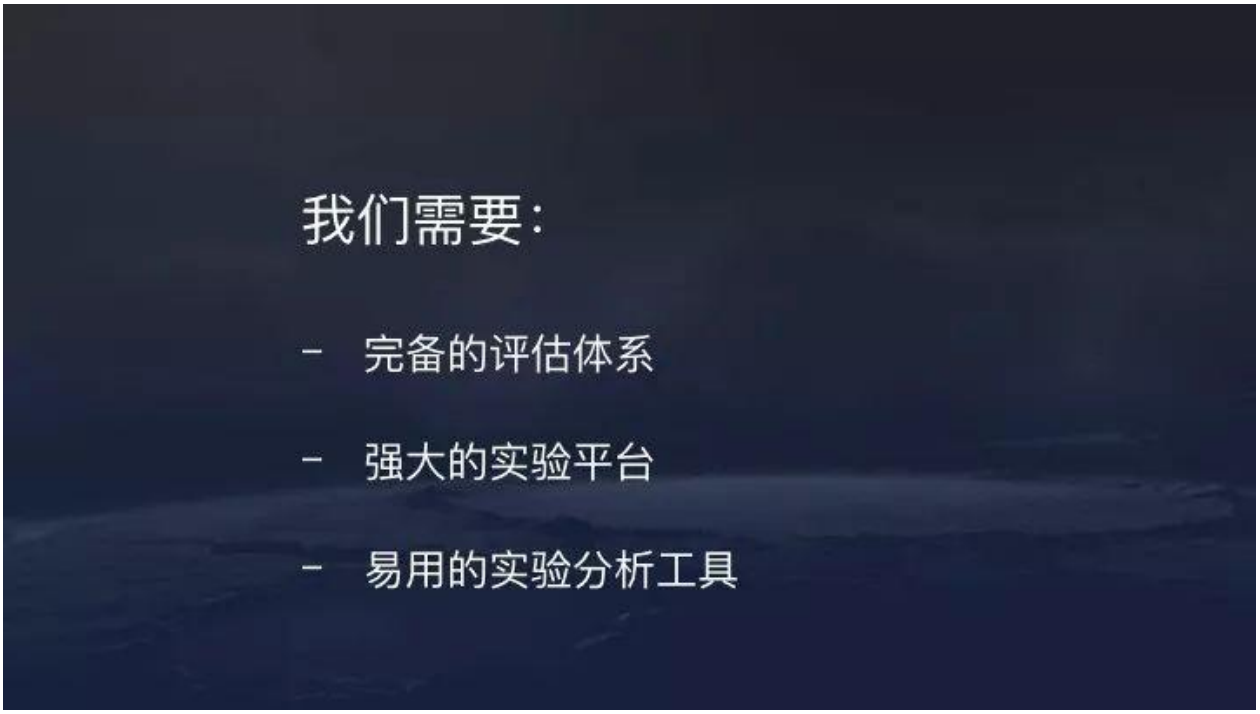
四、评估分析

上面介绍了推荐系统的整体架构，那么如何评估推荐效果好不好？

有一句我认为非常有智慧的话，**“一个事情没法评估就没法优化”**。对推荐系统也是一样。

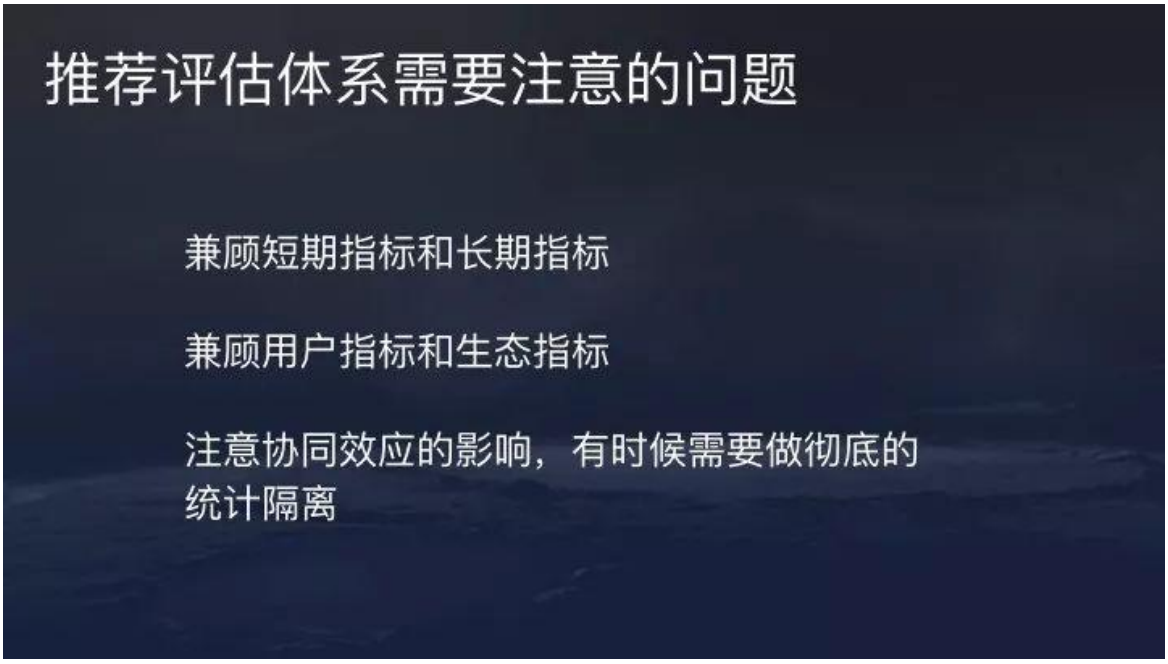


事实上，很多因素都会影响推荐效果。比如侯选集合变化，召回模块的改进或增加，推荐特征的增加，模型架构的改进在，算法参数的优化等等，不一一举例。评估的意义就在于，很多优化最终可能是负向效果，并不是优化上线后效果就会改进。



全面的评估推荐系统，需要完备的评估体系、强大的实验平台以及易用的经验分析工具。所谓完备的体系就是并非单一指标衡量，不能只看点击率或者停留时长等，需要综合评估。过去几年我们一直在尝试，能不能综合尽可能多的指标合成唯一的评估指标，但仍在探索中。目前，我们上线还是要由各业务比较资深的同学组成评审委员会深入讨论后决定。

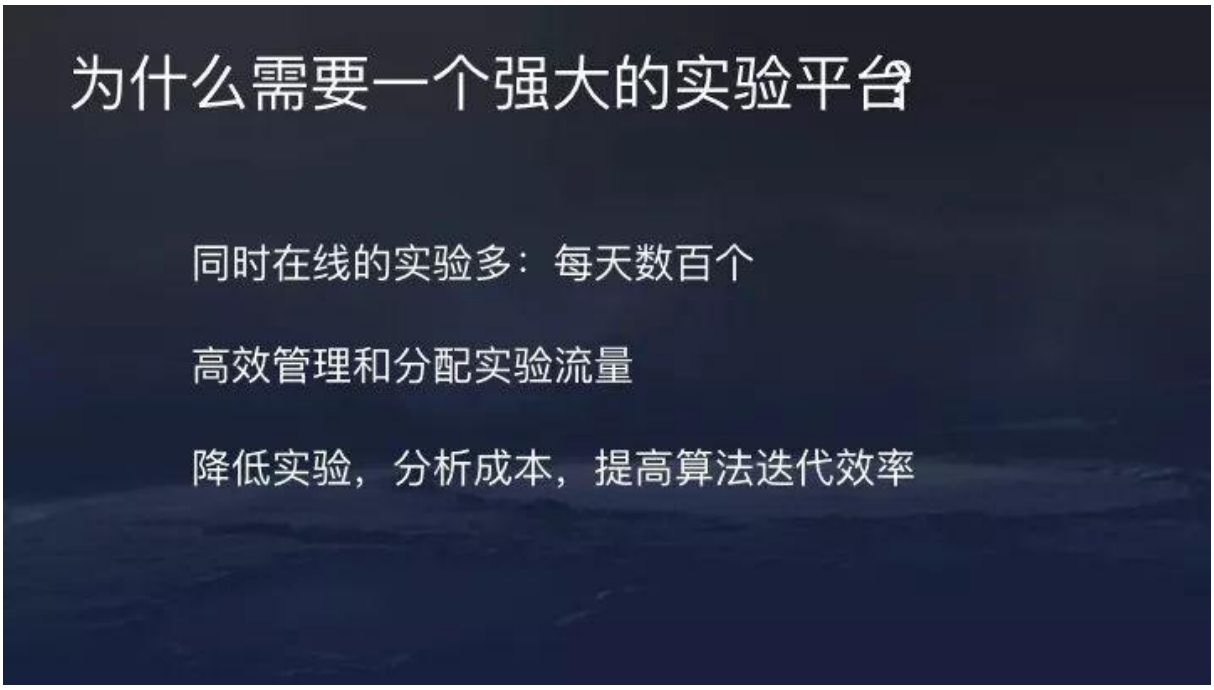
很多公司算法做的不好，并非是工程师能力不够，而是需要一个强大的实验平台，还有便捷的实验分析工具，可以智能分析数据指标的置信度。



一个良好的评估体系建立需要遵循几个原则，首先是兼顾短期指标与长期指标。我在之前公司负责电商方向的时候观察到，很多策略调整短期内用户觉得新鲜，但是长期看其实没有任何助益。

其次，要兼顾用户指标和生态指标。今日头条作为内容分创作平台，既要为内容创作者提供价值，让他更有尊严的创作，也有义务满足用户，这两者要平衡。还有广告主利益也要考虑，这是多方博弈和平衡的过程。

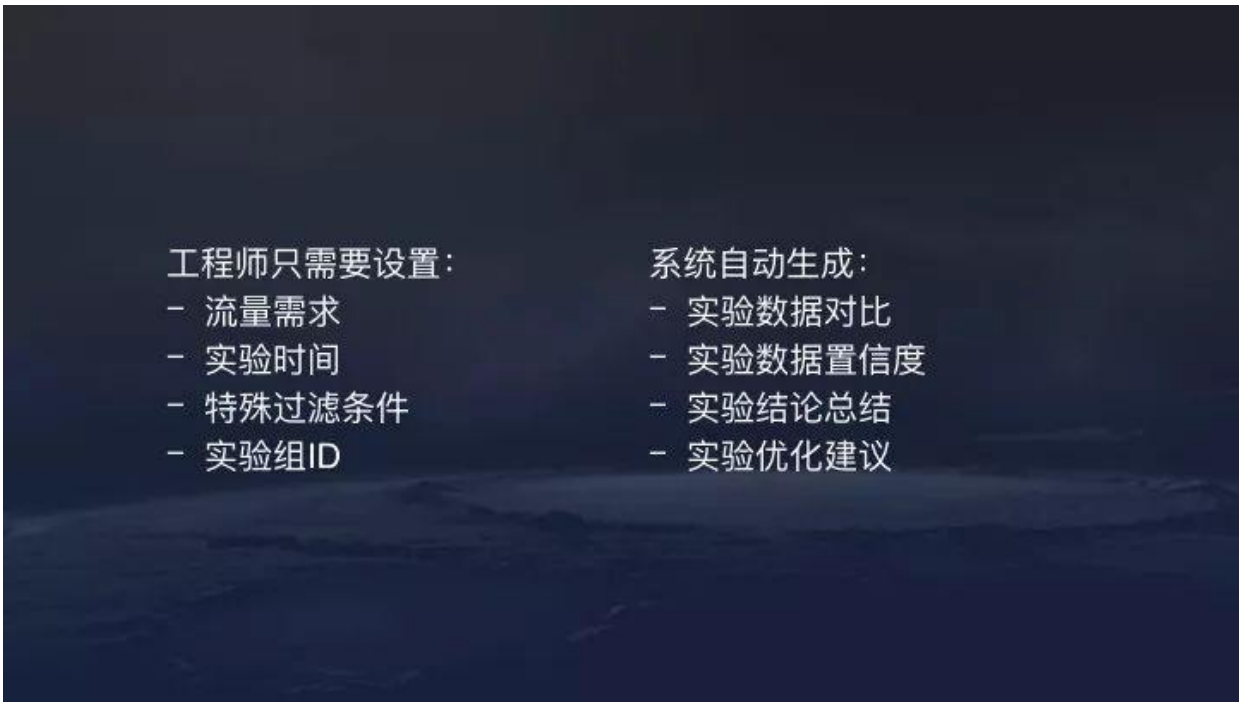
另外，要注意协同效应的影响。实验中严格的流量隔离很难做到，要注意外部效应。



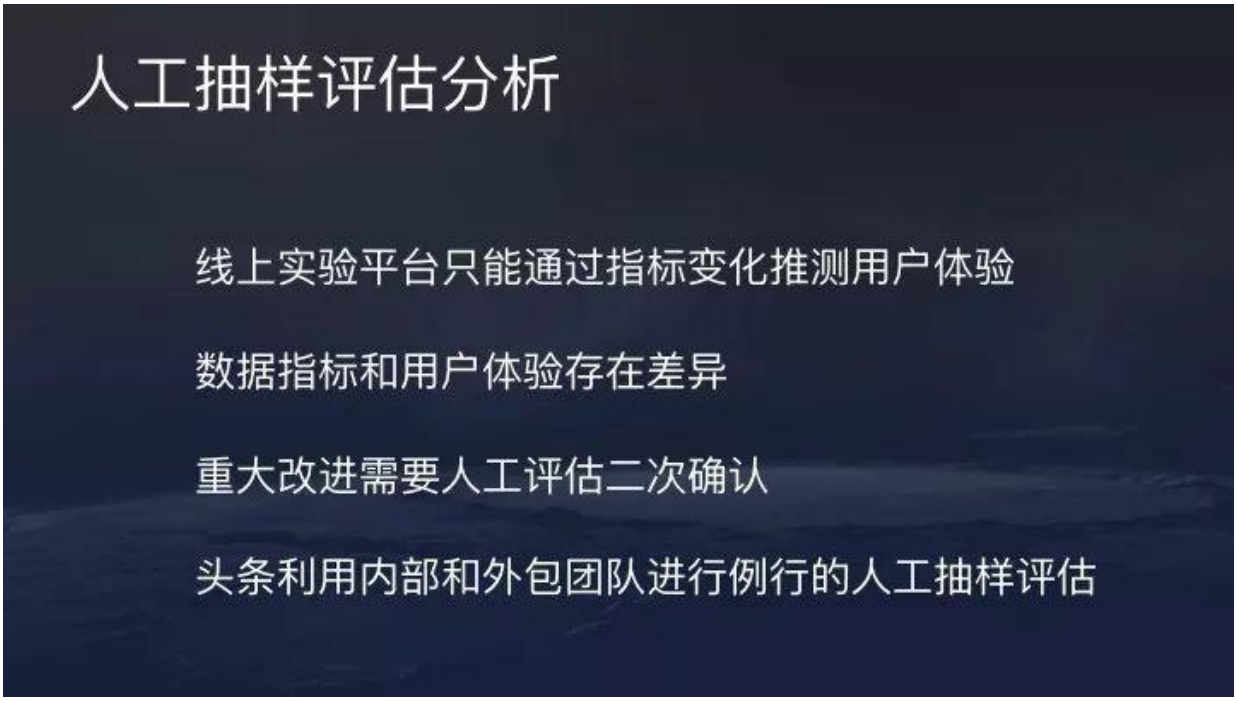
强大的实验平台非常直接的优点是，当同时在线的实验比较多时，可以由平台自动分配流量，无需人工沟通，并且实验结束流量立即回收，提高管理效率。这能帮助公司降低分析成本，加快算法迭代效应，使整个系统的算法优化工作能够快速往前推进。

这是头条 A/B Test 实验系统的基本原理。首先我们会做在离线状态下做好用户分桶，然后线上分配实验流量，将桶里用户打上标签，分给实验组。举个例子，开一个 10%流量的实验，两个实验组各 5%，一个 5%是基线，策略和线上大盘一样，另外一个是新的策略。

实验过程中用户动作会被搜集，基本上是准实时，每小时都可以看到。但因为小时数据有波动，通常是以天为时间节点来看。动作搜集后会有日志处理、分布式统计、写入数据库，非常便捷。

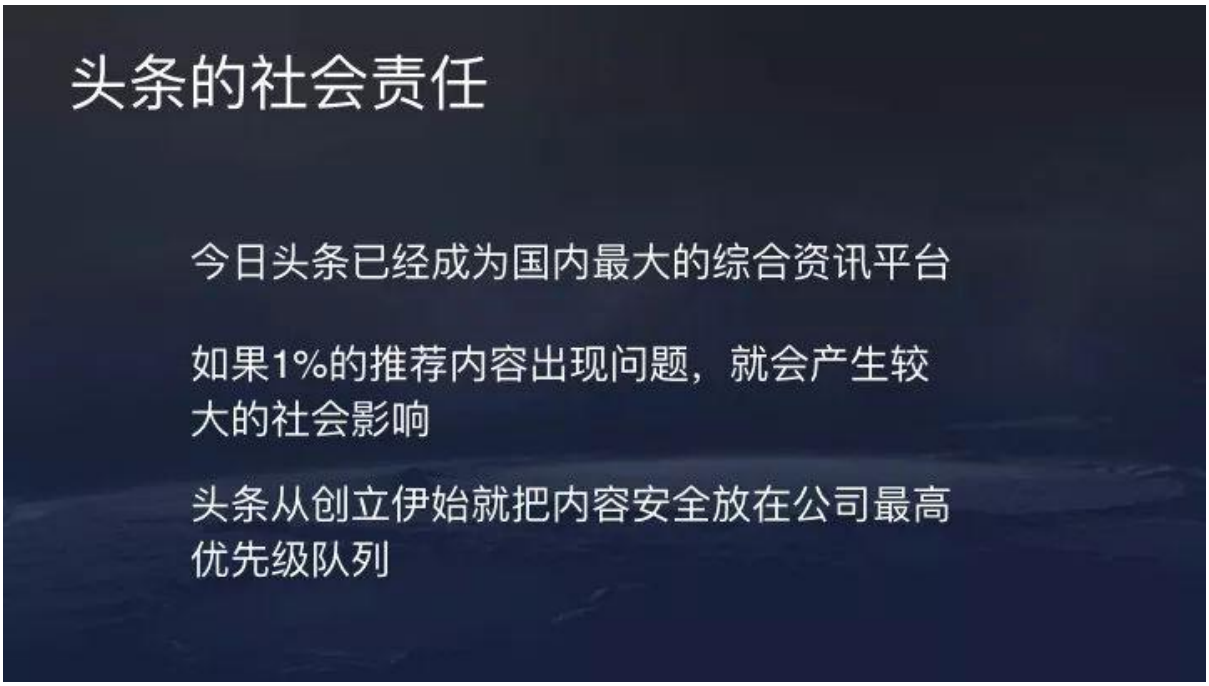


在这个系统下工程师只需要设置流量需求、实验时间、定义特殊过滤条件，自定义实验组 ID。系统可以自动生成：实验数据对比、实验数据置信度、实验结论总结以及实验优化建议。



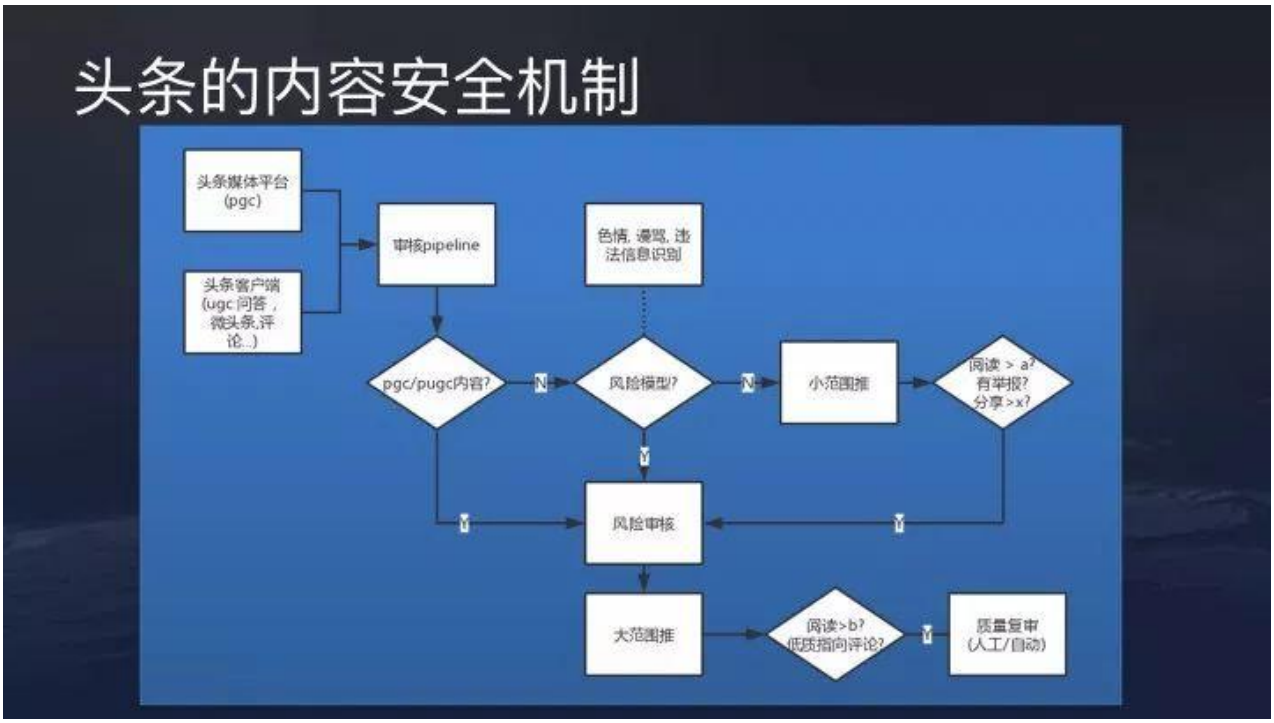
当然，只有实验平台是远远不够的。线上实验平台只能通过数据指标变化推测用户体验的变化，但数据指标和用户体验存在差异，很多指标不能完全量化。很多改进仍然要通过人工分析，重大改进需要人工评估二次确认。

五、内容安全

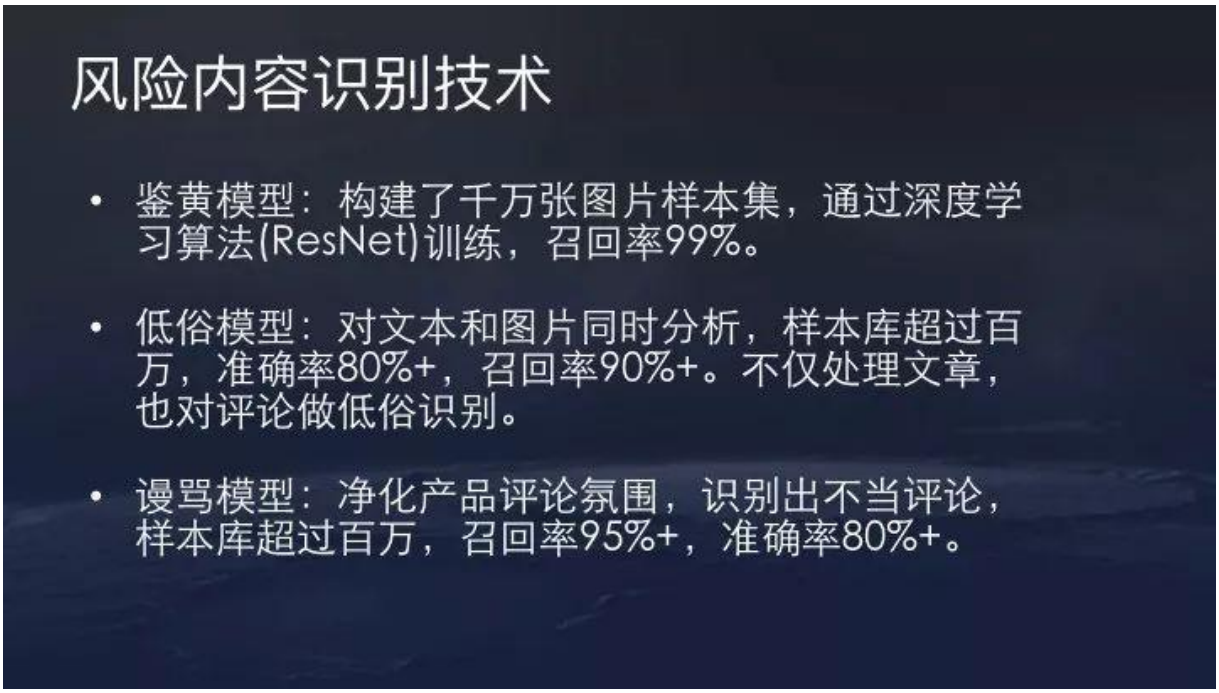


最后要介绍今日头条在内容安全上的一些举措。头条现在已经是国内最大的内容创作与分发凭条，必须越来越重视社会责任和行业领导者的责任。如果 1%的推荐内容出现问题，就会产生较大的影响。

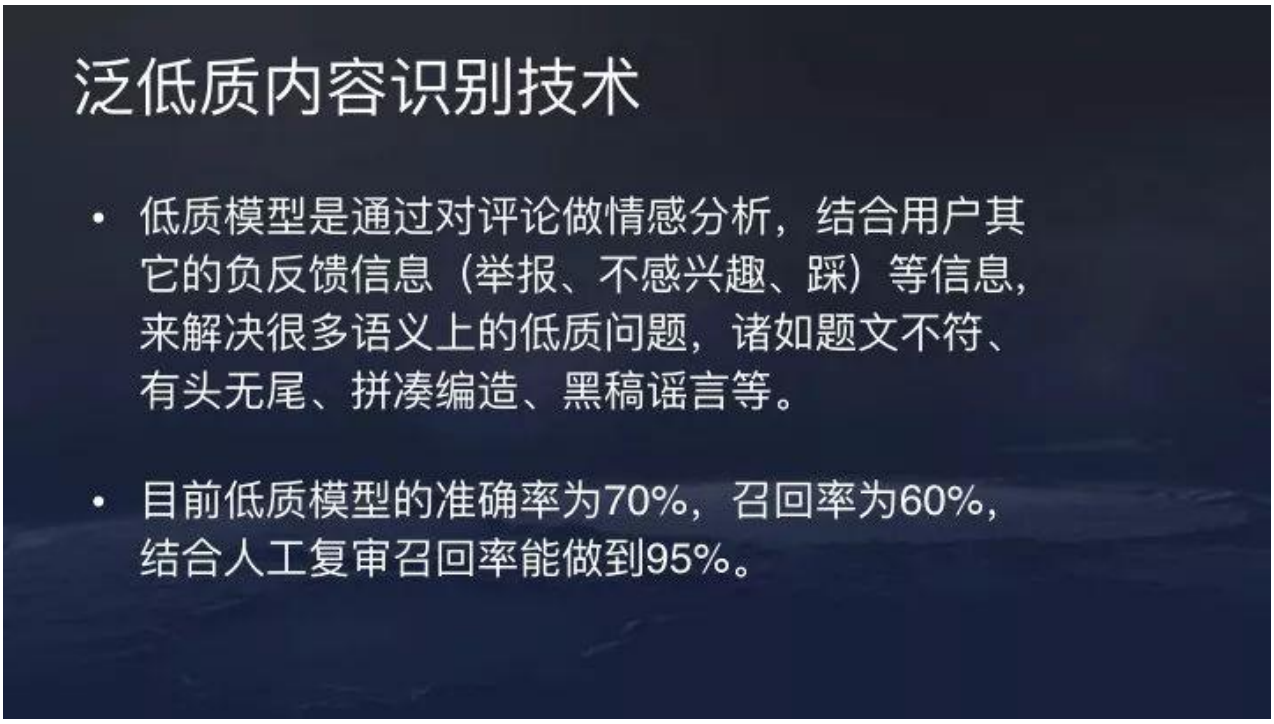
因此头条从创立伊始就把内容安全放在公司最高优先级队列。成立之初，已经专门设有审核团队负责内容安全。当时研发所有客户端、后端、算法的同学一共才不到 40 人，头条非常重视内容审核。



现在，今日头条的内容主要来源于两部分，一是具有成熟内容生产能力的 PGC 平台，一是 UGC 用户内容，如问答、用户评论、微头条。这两部分内容需要通过统一的审核机制。如果是数量相对少的 PGC 内容，会直接进行风险审核，没有问题会大范围推荐。UGC 内容需要经过一个风险模型的过滤，有问题的会进入二次风险审核。审核通过后，内容会被真正进行推荐。这时如果收到一定量以上的评论或者举报负向反馈，还会再回到复审环节，有问题直接下架。整个机制相对而言比较健全，作为行业领先者，在内容安全上，今日头条一直用最高的标准要求自己。



分享内容识别技术主要鉴黄模型，谩骂模型以及低俗模型。今日头条的低俗模型通过深度学习算法训练，样本库非常大，图片、文本同时分析。这部分模型更注重召回率，准确率甚至可以牺牲一些。谩骂模型的样本库同样超过百万，召回率高达 95%+，准确率 80%+。如果用户经常出言不讳或者不当的评论，我们有一些惩罚机制。



泛低质识别涉及的情况非常多，像假新闻、黑稿、题文不符、标题党、内容质量低等等，这部分内容由机器理解是非常难的，需要大量反馈信息，包括其他样本信息比对。目前低质模型的准确率和召回率都不是特别高，还需要结合人工复审，将阈值提高。目前最终的召回已达到 95%，这部分其

实还有非常多的工作可以做。头条人工智能实验室李航老师目前也在和密歇根大学共建科研项目，设立谣言识别平台。

以上是头条推荐系统的原理分享，希望未来得到更多的建议，帮助我们更好改进工作。