ArchSummit
全球架构师峰会 2017

主办方 Geekbang 极客邦科技 InfoQ

人工智能驱动的内容生产与分发

马艳军 博士

Baidu 百度

极客时间

重拾极客精神·提升技术认知

下载极客时间App

获取有声IT新闻、技术产品专栏，每日更新

扫一扫下载极客时间App

主办方 Geekbang. 极客邦科技 InfoQ

AiCon

全球人工智能与机器学习技术大会

助力人工智能落地

2018.1.13 - 1.14　北京国际会议中心

扫描关注大会官网

SPEAKER
INTRODUCE

马艳军

百度 首页业务部主任研发架构师

马艳军博士，现在在百度主要从事自然语言处理、机器翻译、推荐技术的研发工作，相关研发成果广泛应用于百度搜索、信息流等产品中。
马博士曾参与欧盟第七框架计划（FP7）、863计划等国内外项目，发表论文20余篇，申请国内外技术发明专利20余项，并多次担任ACL，IJCAI等国际权威会议领域主席和审稿人，相关成果曾获得2016年国家科技进步二等奖。

提纲

- 内容消费的行业现状
- 人工智能辅助内容生产
- 内容质量控制模型
- 内容分发/推荐系统

- 内容消费的行业现状

什么是内容消费？

一种直接或间接以内容产品和内容服务为消费对象的经济活动

媒介独立

内容传播的媒介相对独立，传播媒介的属性限定了内容的类型，如视频类内容几乎只能在电视频道上传播，供需调节难

角色单一

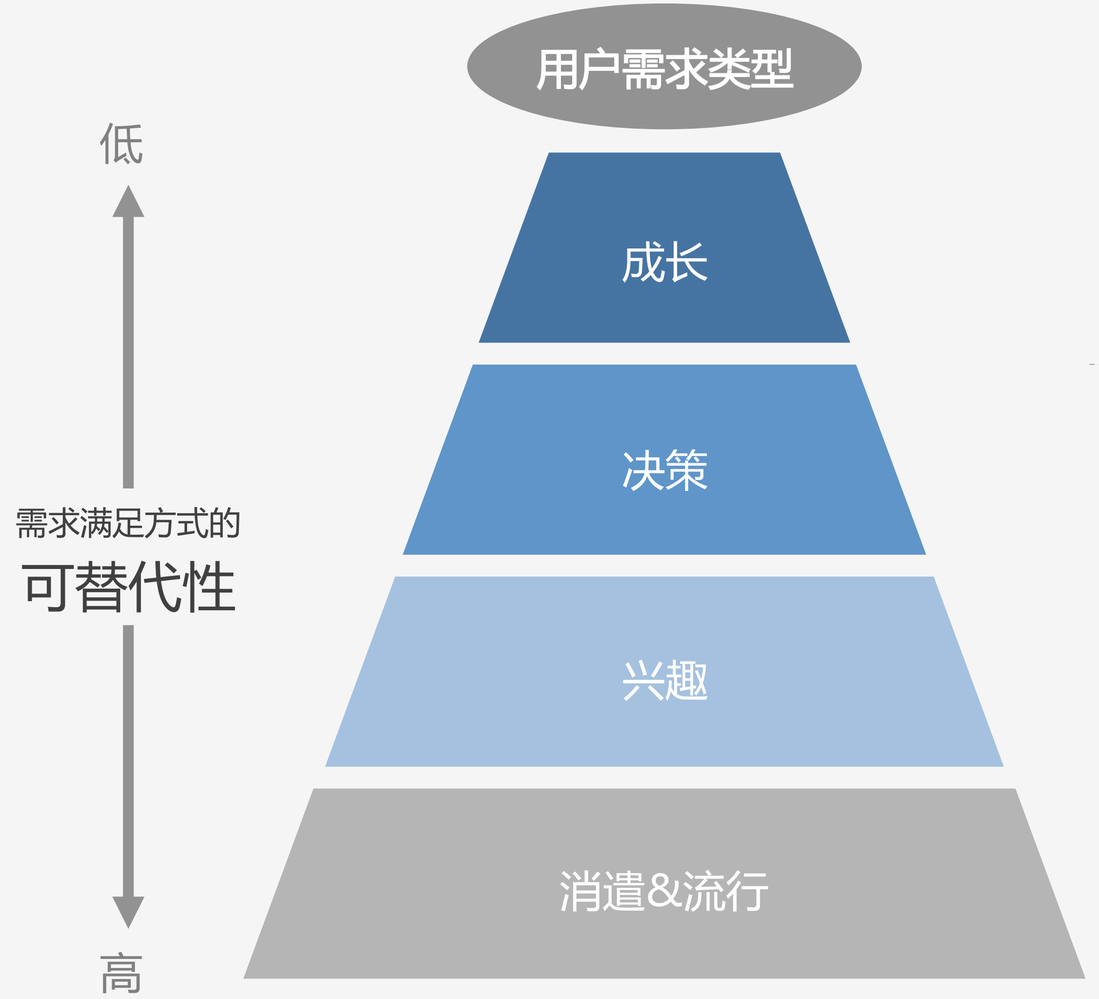
内容生产成本高；图文内容、视频内容的生产者之间有较大的壁垒，内容的供给以生产者为主，内容产量相对低

单向消费

内容从生产者、传播渠道、到用户，全程单向传播，一次内容消费的完成就意味着一次传播行为的终结

内容消费满足了用户的什么需求？

覆盖每一个用户生活的方方面面

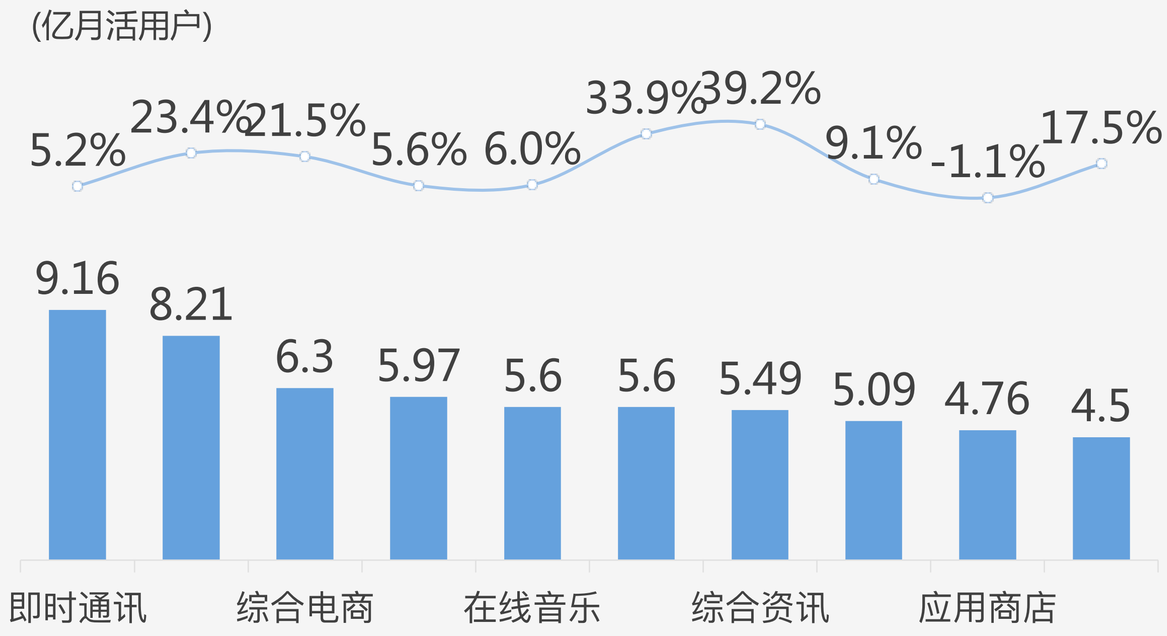


需求特点	内容供给
刚需 · 少量	观点型内容
刚需 · 适量 · 粘性	知识型内容
高频 · 海量 · 碎片	泛娱乐内容

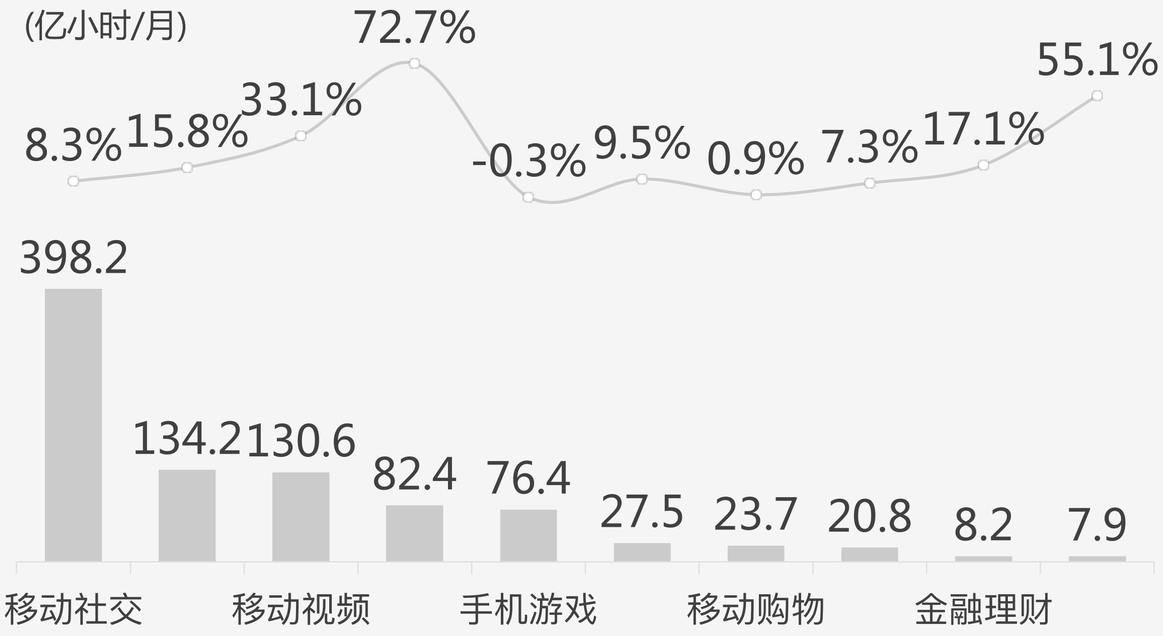
内容消费为什么重要？

1. 内容消费已经发展成为了一个影响所有人日常生活的庞大产业

Top 10行业用户规模及同比增长率（2017年6月）



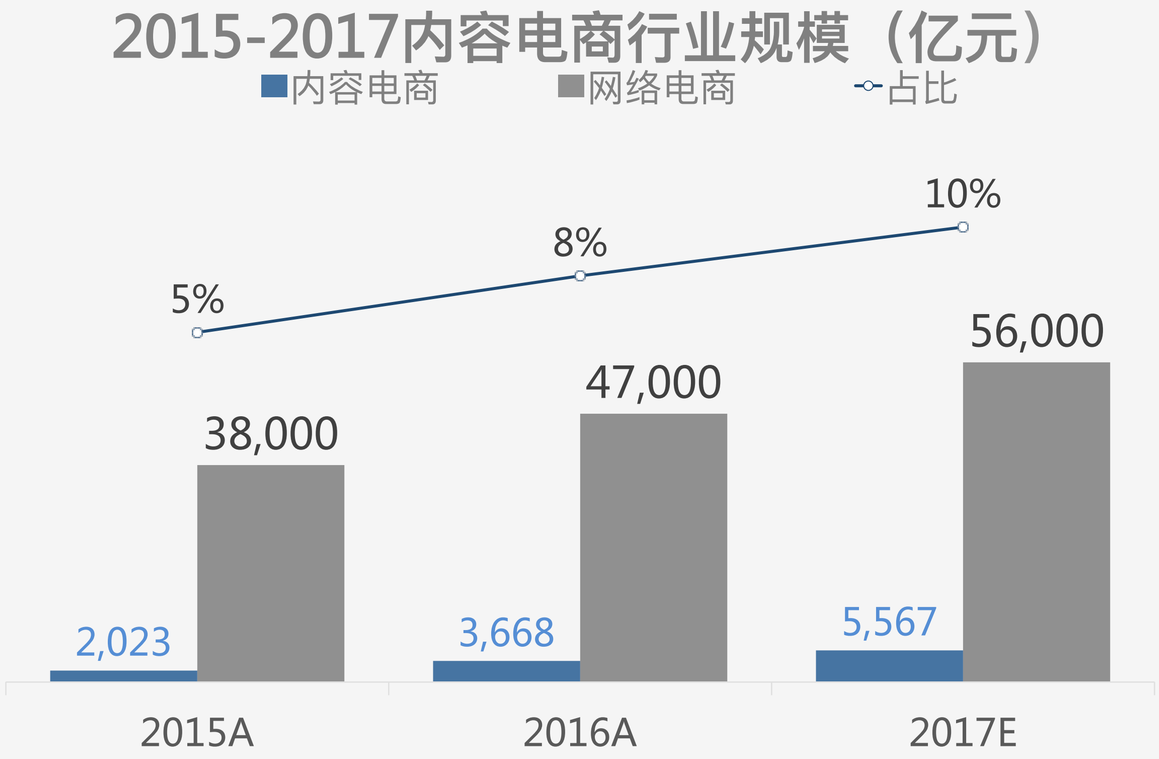
Top 10行业用户总时长及同比增长率（2017年6月）

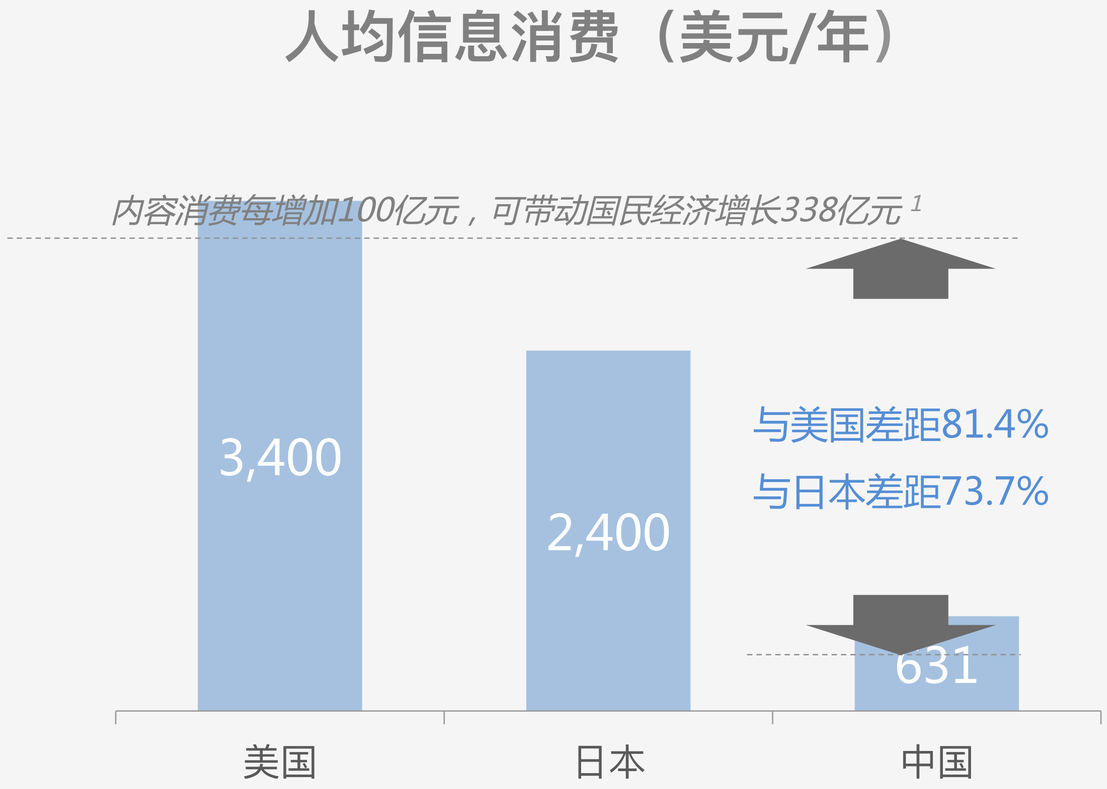


数据来源：Top10行业用户规模、用户总时长和同比增长率来源于Quest Mobile的《移动互联网2017年Q2夏季报告》

内容消费为什么重要？

2. 更是国民经济杠杆，关乎社会主流价值观的塑造





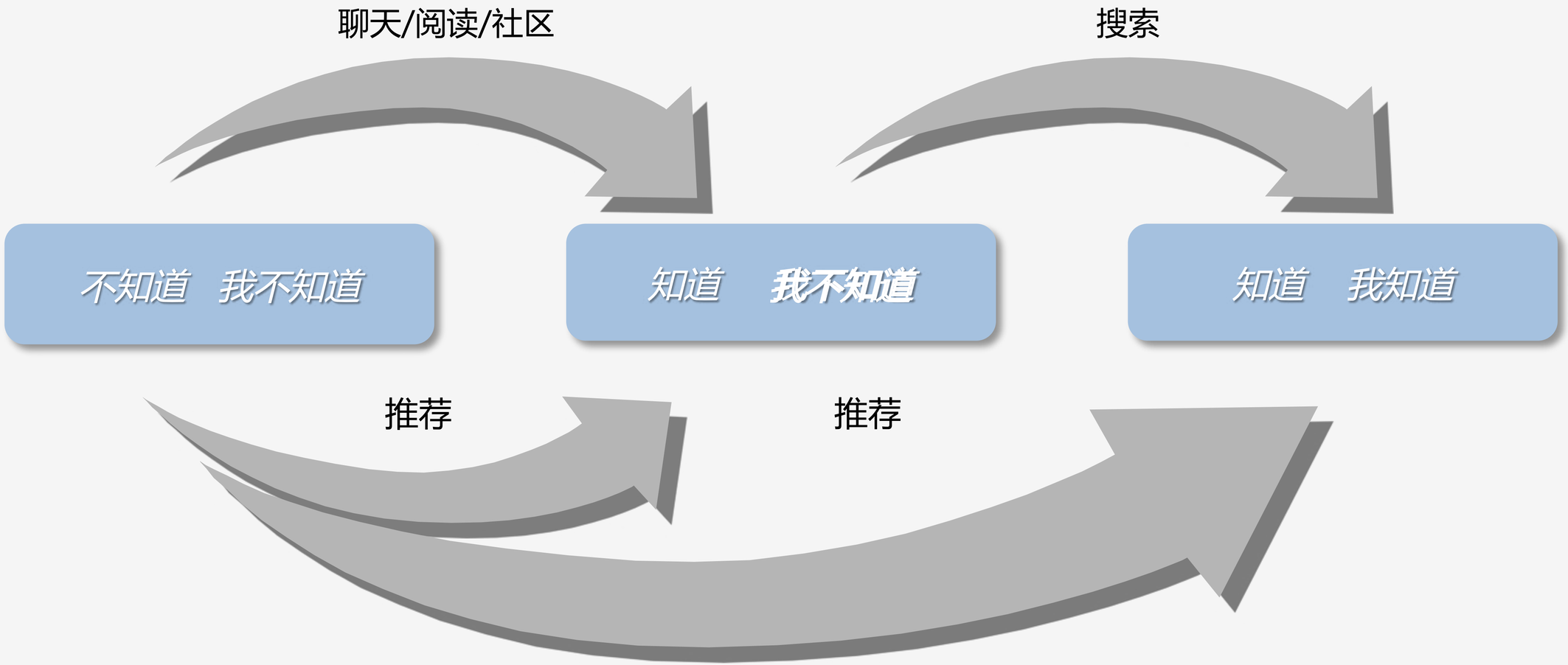
数据来源：1）2013年工业和信息化部电信研究院数据；2）人均信息消费：世界银行及恒丰银行研究院商业银行研究中心统计数据

内容消费行业地图

人工+机器，多种消费形态组合

内容消费的主要形态

从搜索到推荐，是内容获取方式的必然演进



内容消费的三大变化趋势

媒介融合 · 角色融合 · 消费升级

媒介融合

- 一个媒介上可以获取多种内容，同一内容也可以在多个媒介上分发
- 获取内容的方式大发展，从传统的人找内容，发展到内容找人

角色融合

- 生产门槛降低，人人都是生产者，人人都是消费者，内容生产者猛增
- 从专业生产（PGC）发展到社会化生产（UGC），内容海量爆发

消费升级

- 从关注到最终的分享可循环，内容散落在消费的各个环节，影响力扩大
- 消费结构变化，用户需要更多专业垂直、精细、多元化的内容

需要解决的问题

- 作者深度赋能
- 优质内容识别
- 精准个性推荐

百度智能驱动的内容消费生态

始于2016年6月

定制生产、优质生产、高效生产

- 选题推荐：大数据+知识图谱
- 内容选材：知识图谱+自动配图
- 图片处理：智能裁图
- 文字处理：自动标题推荐+语义纠错
- 视频识别：视频去重+长短关联
- 智能写作：AI智能写作

精准个性化推荐及推送

- 搜索：从关键字，到语音、图片交互
- 个性化推荐：百度特有的超大规模实时个性化推荐系统，千亿规模参数下的多目标最优化
- 推送：基于用户兴趣点、地理位置、天气等综合因素

多维度识别优质内容资源

- 质量先验：基于半监督学习构建对于篇章质量的全方位的理解框架，多维度丰富信息
- 质量后验：基于NLP和CNN，识别用户正负反馈，评价内容质量

搜索+推荐，满足用户内容需求

双引擎

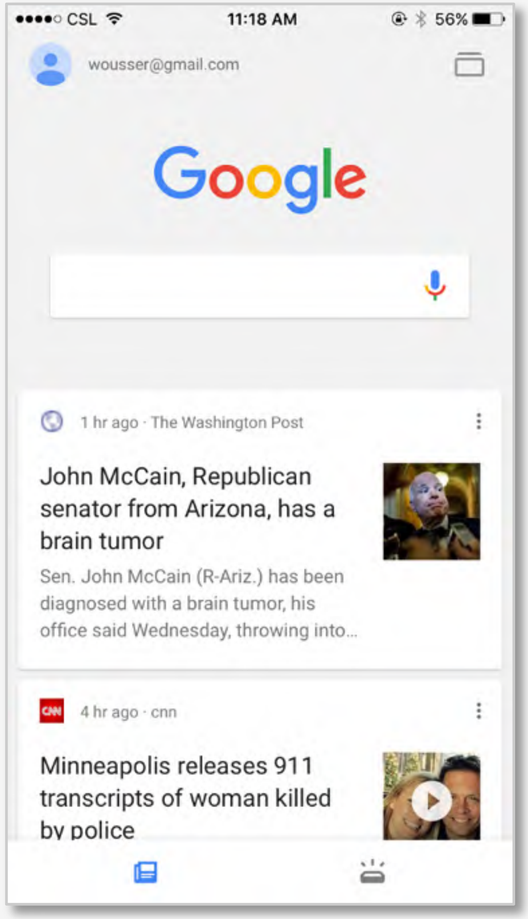
手机百度，“搜索+推荐”双引擎内容分发

极速搜索

一站搜索全网，搜文字、搜图片，快速识别，帮助用户主动获取所需内容

精准推荐

精选资讯、视频、小说、漫画等优质内容，结合智能个性化推荐算法，满足用户潜在内容消费需求



Google Feed, 2017年7月上线

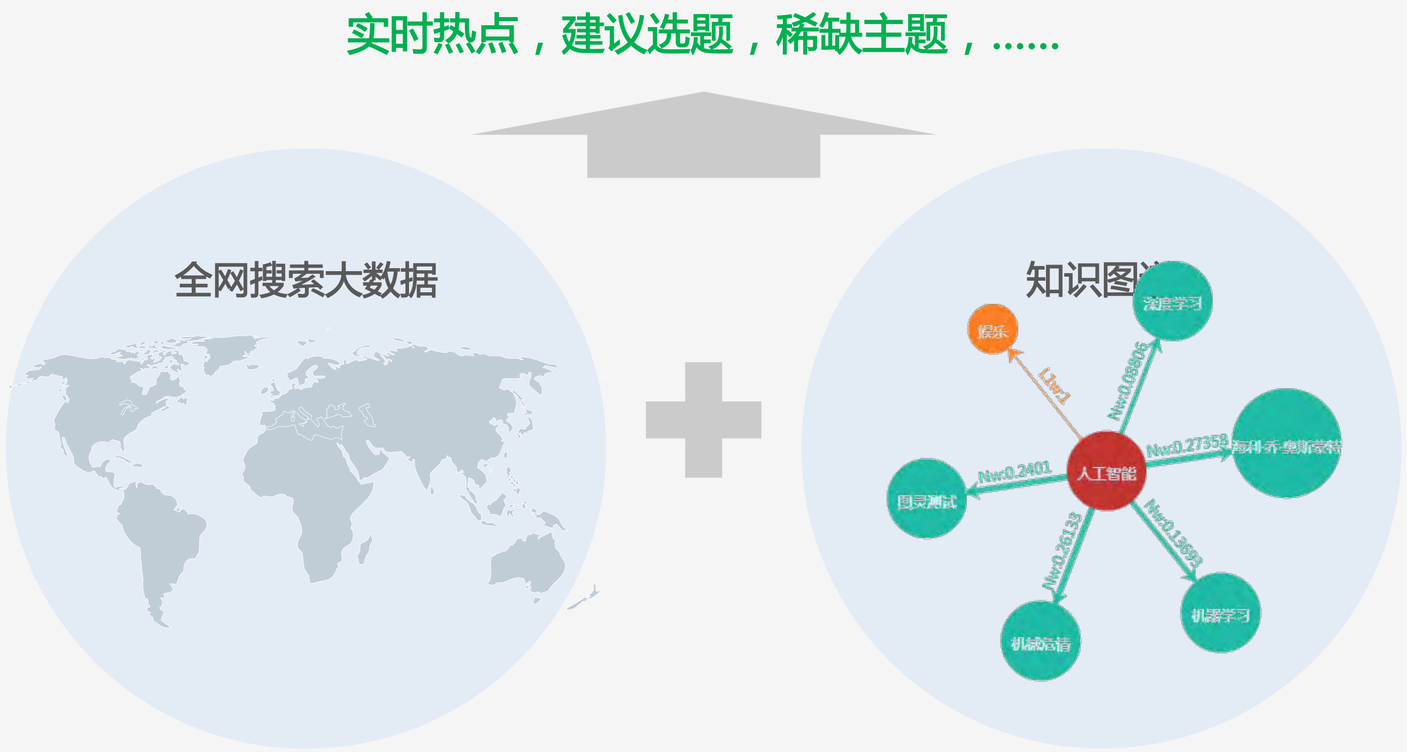
- 人工智能辅助内容生产

内容生产流程 1

选题推荐 · 内容选材 · 图片处理 · 文字处理 · 视频识别 · 智能写作

选题推荐

- 基于7亿互联网民每天60亿次搜索请求的大数据积累，挖掘实时热点事件，结合用户画像和知识图谱技术，智能推荐给相关作者，邀请作者写作，解决作者选题环节的痛点
- 相比普通作者自选题文章平均用户点击率提高2.7倍



内容生产流程 2

选题推荐 · 内容选材 · 图片处理 · 文字处理 · 视频识别 · 智能写作

内容选材

- 基于知识图谱的内容推荐：通过语义分析技术识别作者写作主题，结合百度知识图谱、指数、百科、正版图库等产品，帮助作者发现写作素材

自动配图

- 基于图像识别能力的自动配图功能，根据图像清晰度、美观度、图文相关性等特征选取优质图片
- 模糊图片识别准确率71%，清晰图片识别准确率96%以上；自动配图准确率90%

内容生产流程 3

选题推荐 · 内容选材 · 图片处理 · 文字处理 · 视频识别 · 智能写作

智能裁图

- 利用基于深度学习的图像处理技术，自动识别图片主要内容区域并进行截取
- 准确率97%以上，百家号文章头图的不合格率降低60%

内容生产流程 4

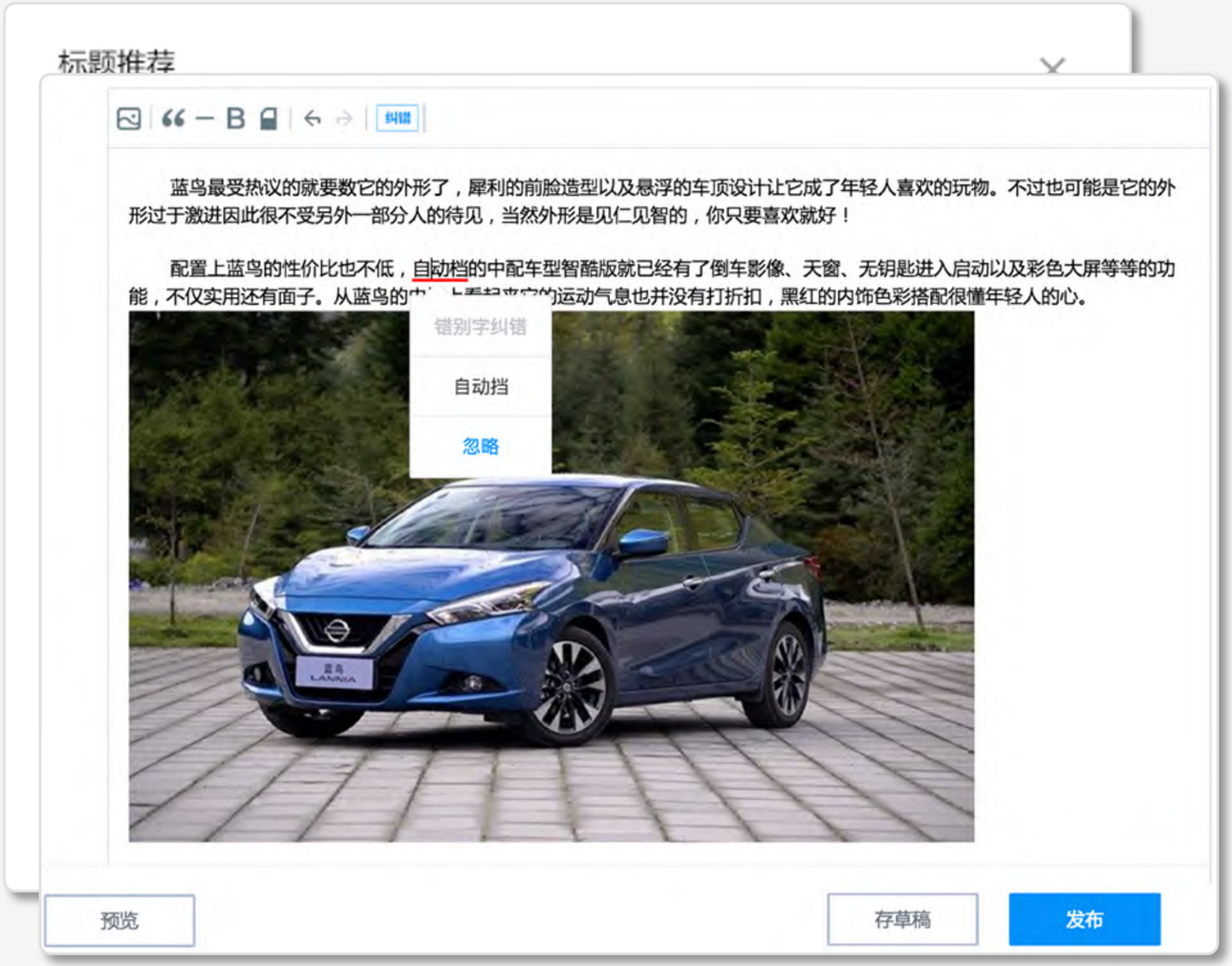
选题推荐 · 内容选材 · 图片处理 · 文字处理 · 视频识别 · 智能写作

标题推荐

- 基于语义理解技术，识别文章内容，推荐优质标题

语义纠错

- 运用机器学习的能力，让机器理解文章中的句子含义，来找出不符合语义的错别字
- 纠错准确率99%



内容生产流程 5

选题推荐 · 内容选材 · 图片处理 · 文字处理 · 视频识别 · 智能写作

视频识别

- 视频查重：基于图像及语音的比对技术，识别内容近似的短视频，应用于审核前原创作者保护，线上降低重复视频推荐；准确率达到97.5%

- 长-短视频关联：将影视截取的短视频匹配到其来源长视频，利用知识图谱获取关联信息、进行相关推荐；准确率90%

--短视频片段获取长视频结构化信息--

影视剧名称：三生三世十里桃花

演员：杨幂，赵又廷，张智尧、迪丽热巴

年份：2017

类型：古装剧

内容生产流程 6

选题推荐 · 内容选材 · 图片处理 · 文字处理 · 视频识别 · 智能写作

智能写作

- 大数据+知识图谱，自动挖掘现实生活中动态热度变化，再通过算法自动生成文章

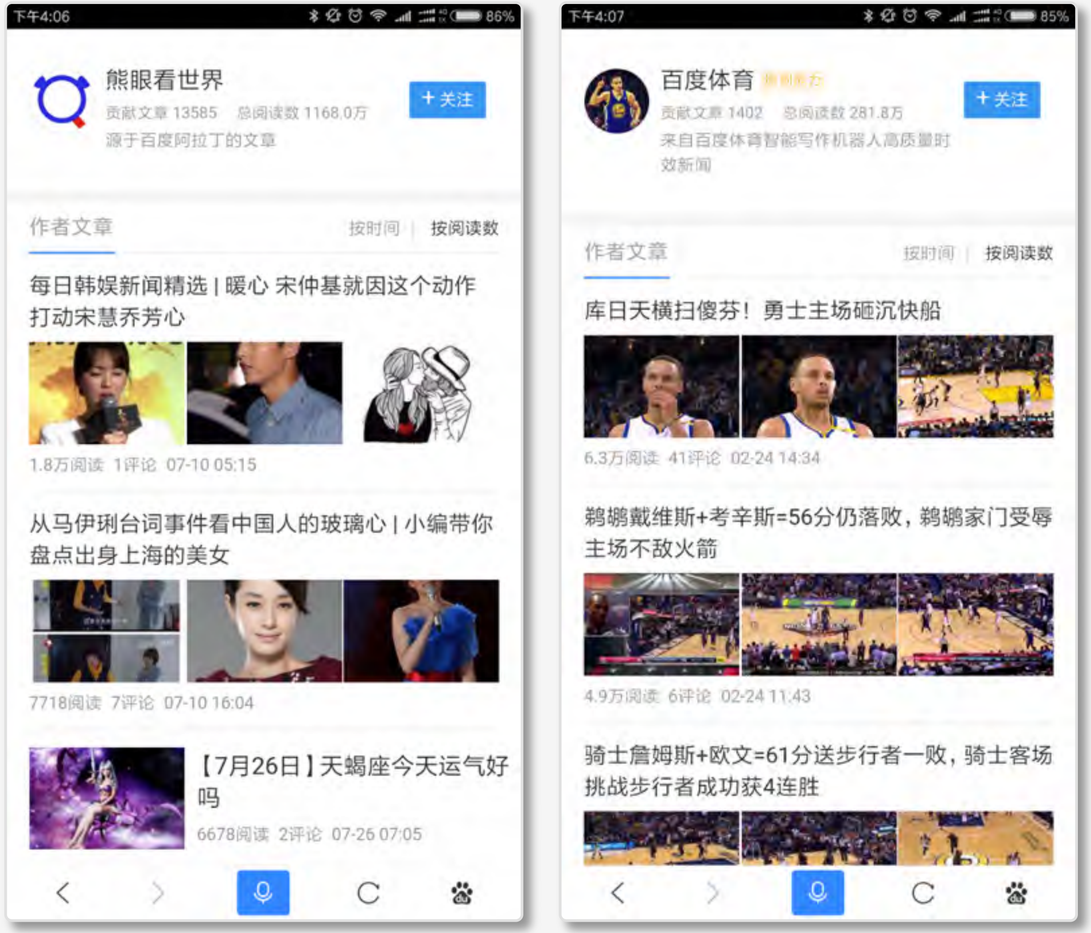
2014年美联社就启用机器人进行财经报道，每季度能完成对4000家公司的财报报道，此前靠记者仅能完成400家

2015年起，机器人用于体育新闻报道，完成梦幻橄榄球赛报道，还在行文中大展幽默（技术支持为自动化洞察力公司，其开发的软件Wordsmith在16年共写了15亿篇文章，涉及金融、房地产、体育等多领域）

2015年9月，推出自动化新闻写作机器人Dreamwriter，发布了关于消费价格指数的报道

2015年11月启用机器人写稿系统“快笔小新”，供职于体育部、经济信息部和中国证券报，撰写体育和财经稿件

百度智能写作示例



- 内容质量控制模型

内容质量控制模型

构建全方位的篇章质量理解框架，多维度理解内容质量

低质内容示例

篇章质量理解框架

低质内容识别

- 内容低俗
- 内容令人不适
- 标题党
- 旧闻新发
- 恶意广告
- 基础文法硬伤

内容优质度计算

- 可读性
- 时效性
- 原创度
- 源权威度
- 内容影响力
- 题材吸引力

基础数据

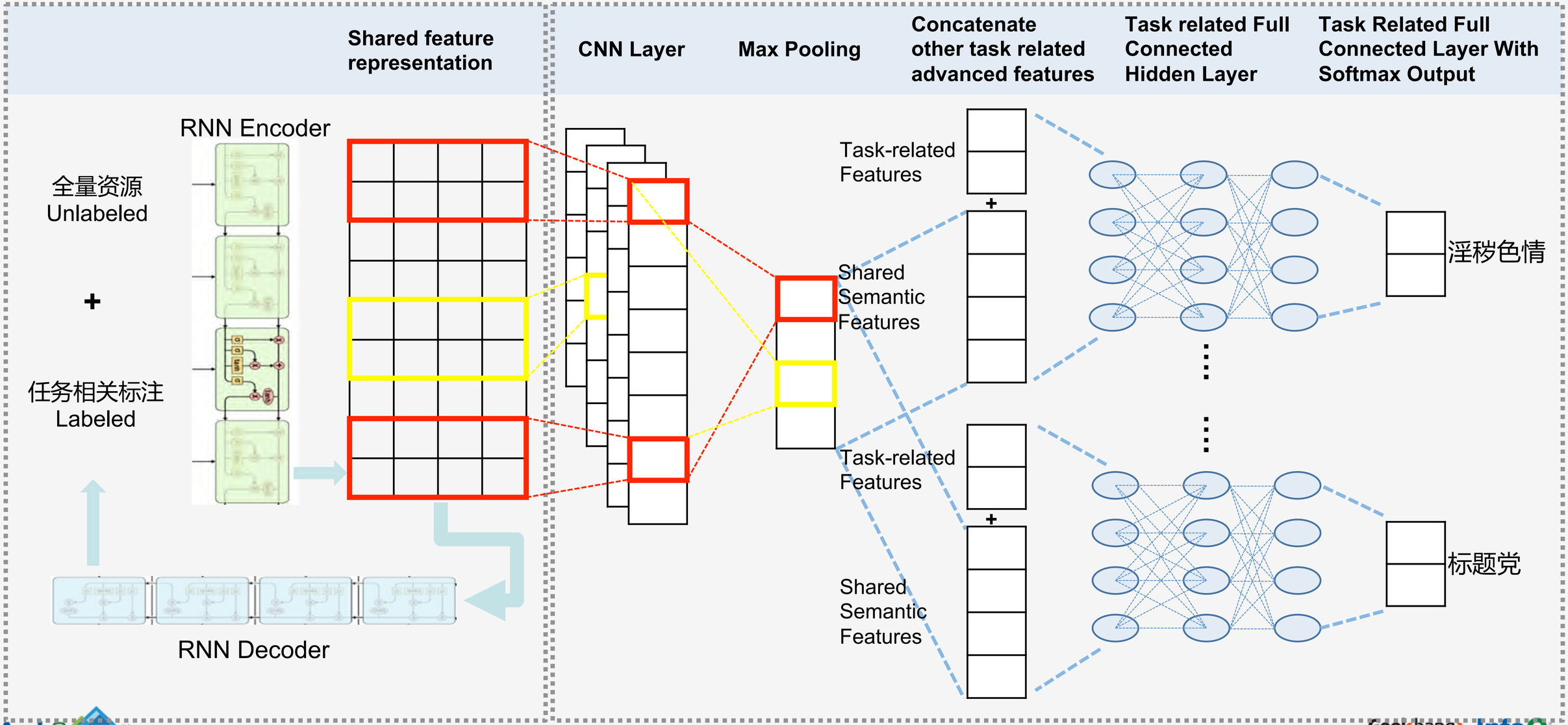
- 作者行为
- 全网用户行为
- 作者信息
- Feed用户行为

基础技术

- 网页搜索
- 机器学习
- 自然语言理解
- 数据挖掘

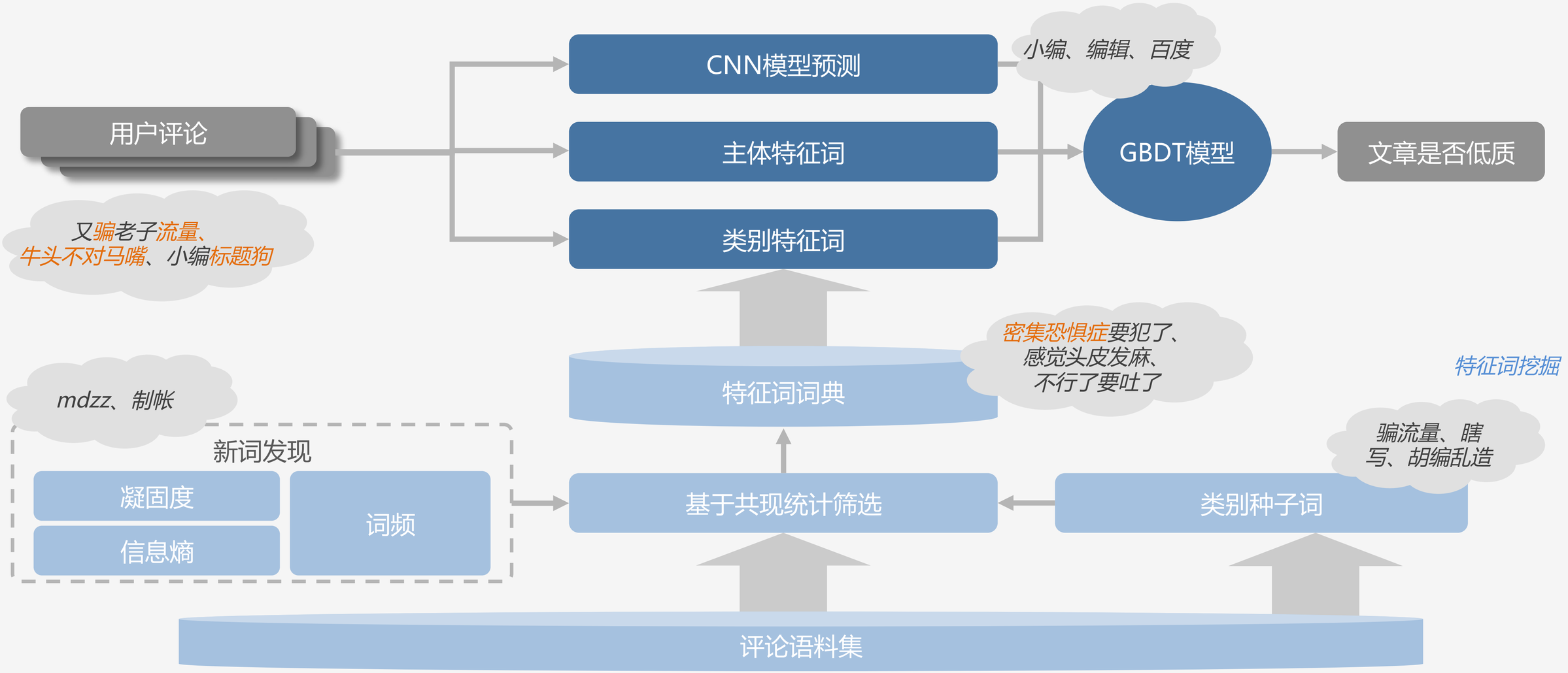
内容质量控制模型

先验质量识别：基于半监督的多目标统一内容分类框架



内容质量控制模型

后验质量识别：基于NLP语义理解的评论正负反馈分析

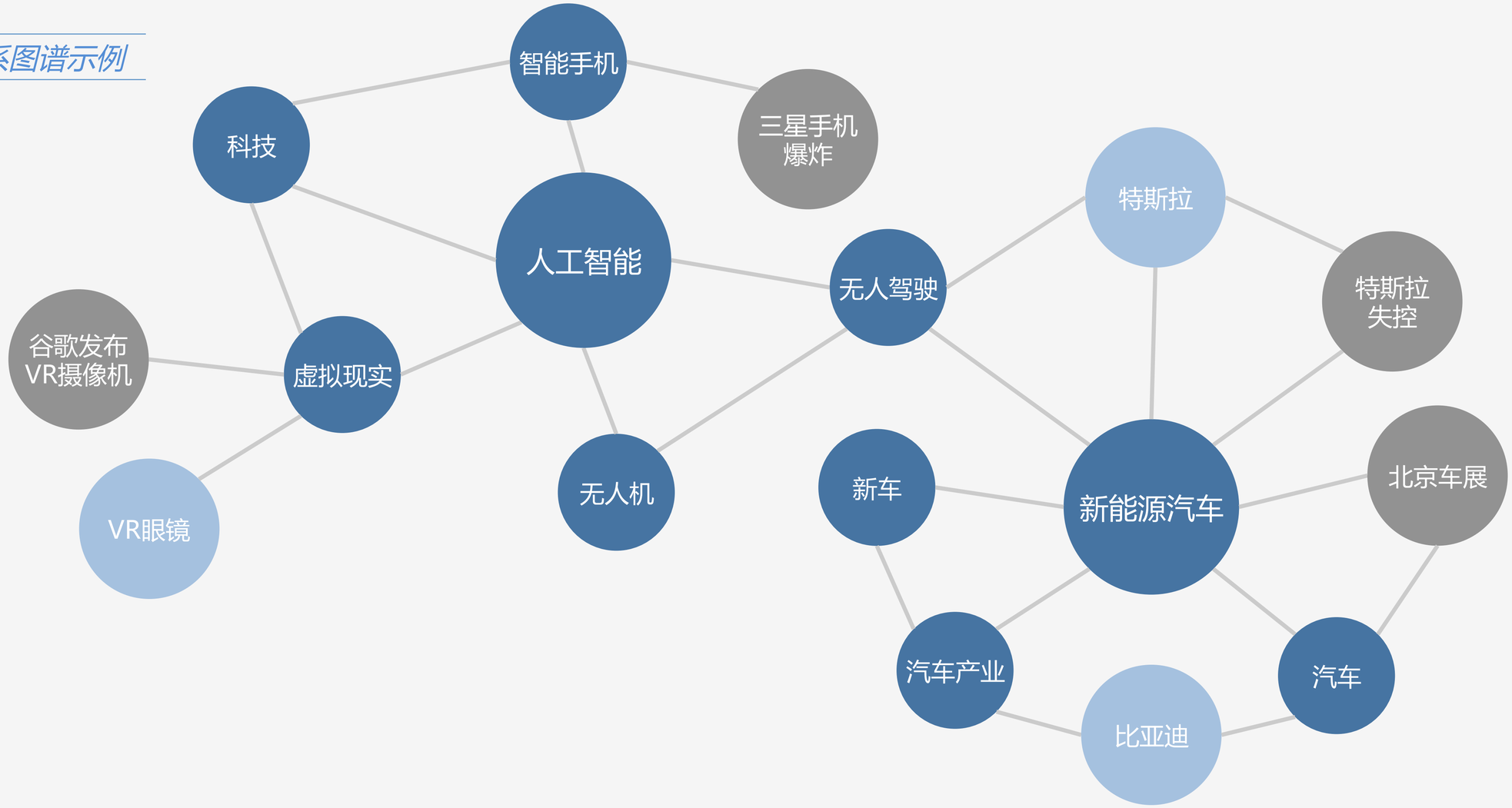


- 内容分发/推荐系统

内容分发/推荐系统

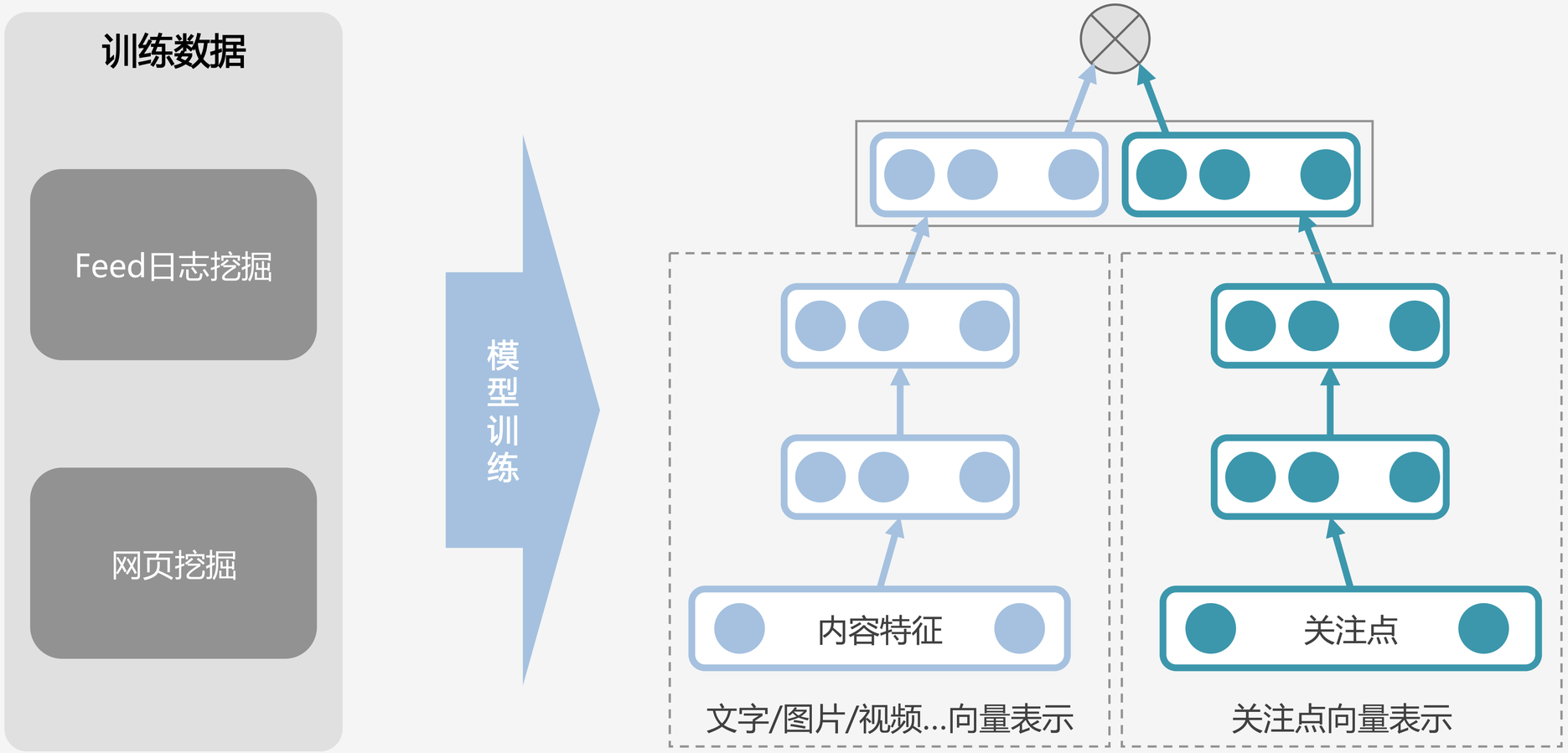
基于知识图谱，构建关注点关系图谱

关注关系图谱示例



内容分发/推荐系统

内容理解：关注点抽取

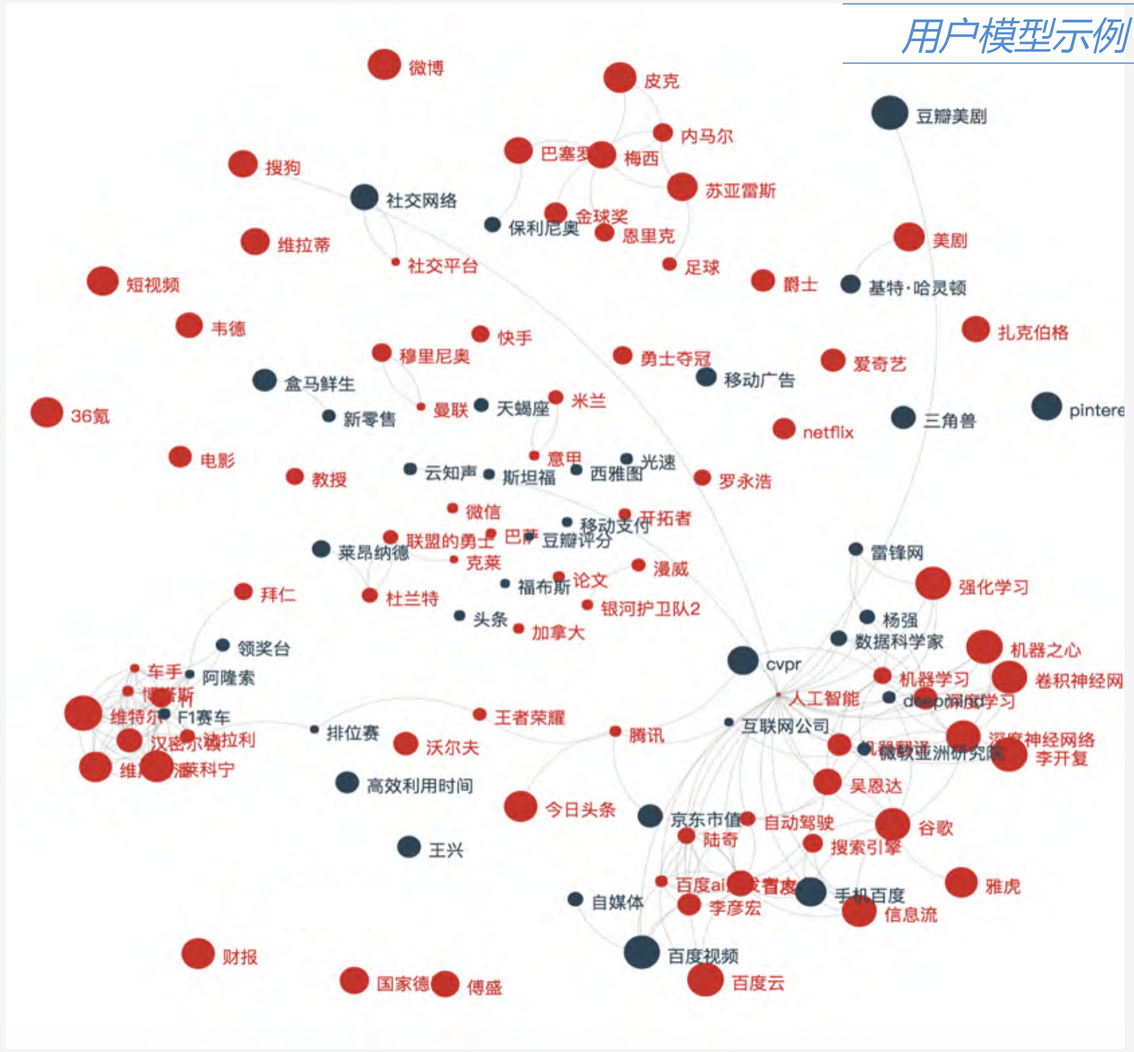


内容分发/推荐系统

用户理解：用户兴趣挖掘

- 数据来源
  - 行为数据：搜索query、Feed阅读、贴吧、全网浏览数
  - 画像数据：百度用户画像数据 ....
- 基于行为数据的挖掘策略
  - 针对搜索Query和Feed阅读内容，基于内容理解技术提取兴趣主题和关注点
  - 根据关注点/主题在内容中的权值、展现数、点击数计算相应的强度
  - 根据点击数和持续周期，区分长期和短期兴趣点
- 基于画像数据的挖掘策略
  - 画像兴趣体系到Feed兴趣体系映射
  - 基于画像属性的人群划分和推荐触发

用户模型示例

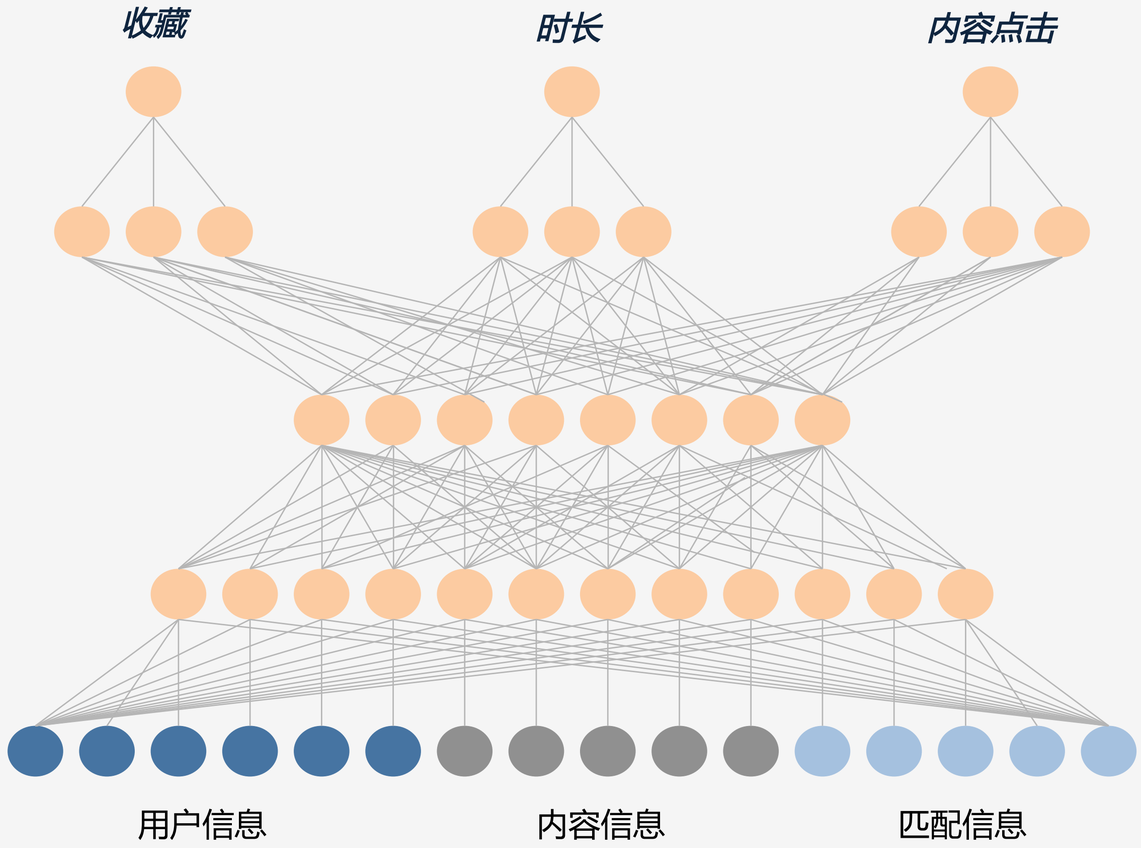


内容分发/推荐系统

推荐策略：多维度的推荐队列召回、排序和融合

推荐价值相关的多目标学习系统（Multi-Task Learning

- 模型输入：千亿规模参数
  - ➢ 用户信息：兴趣、属性、偏好，等
  - ➢ 内容信息：吸引力、质量、相关性，等
  - ➢ 匹配信息：用户-内容
- 模型输出：多目标
  - ➢ 内容点击、时长、收藏、评论、分享，等



内容分发/推荐系统

DNN上线，核心兴趣点强相关内容推荐准确度显著提升

DNN上线前

用户核心兴趣点

DNN上线后

THANK YOU

如有需求，欢迎至 [讲师交流会议室] 与我们的讲师进一步交流

ArchSummit
全 球 架 构 师 峰 会 2017