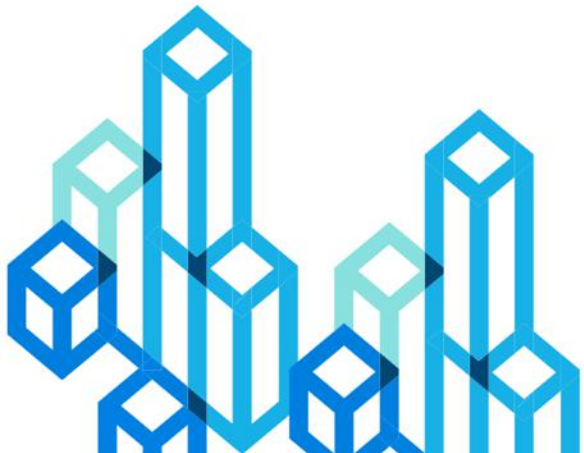
趣头条推荐系统用户画像构建



个人简介

- 2015年7月研究生毕业于西安交通大学
- 先后就职于百度、小米、趣头条
- 在NLP、知识图谱、智能问答、用户画像等方向有相关项目经验
- 目前就职于趣头条，负责AI Lab团队日常工作

趣头条介绍

趣头条是一款以娱乐、生活资讯为主体内容，依托于智能化数据分析系统，为新兴市场受众提供精准的内容分发服务的APP

目录

目录

什么是用户画像

年龄段要求： 1988-1995，不要94年

身高要求： 176-188

外型偏好： 肤白，高瘦，单眼皮

学历要求： 一本本科及以上

现居地要求： 武汉

籍贯要求： 二线城市 独生子

是否接受烟酒： 可以接受

是否介意恋爱史/婚史： 不要离异

职业偏好： 不要销售；受过良好的教育，对自己的工作和未来有切实规划。

收入要求： 年薪12w+

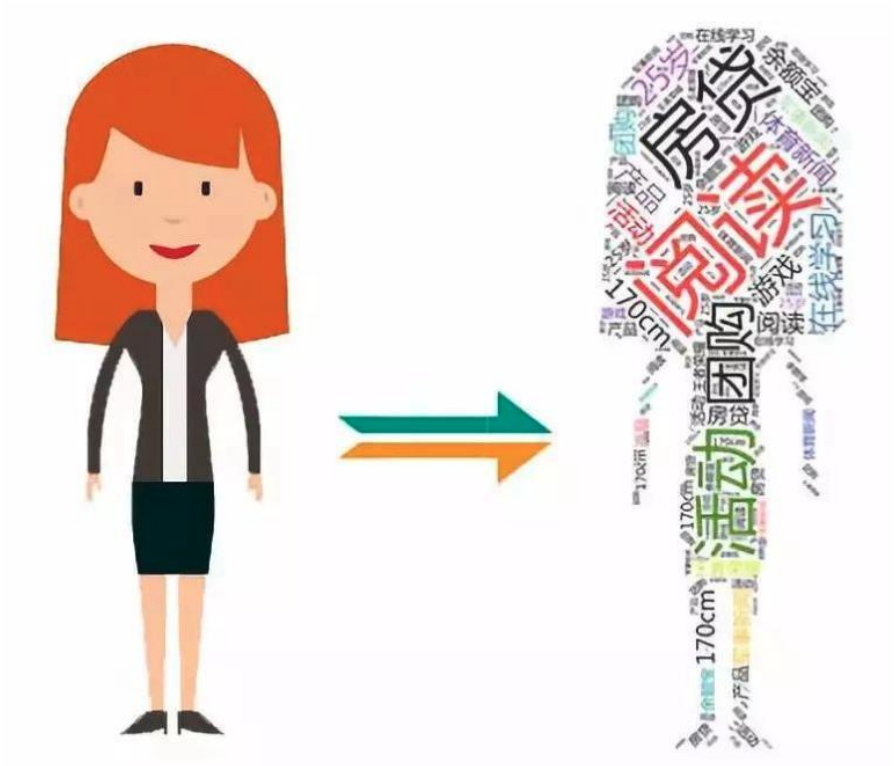
住房购车要求： 有房

家庭背景要求： 家庭本分厚道

偏好性格： 人品好，有责任，有担当，有爱心，喜欢小动物。

不能接受的点： 喜欢打牌，花心，没有责任心

什么是用户画像



Alan Cooper（交互设计之父）最早提出了 persona 的概念:“Personas are a concrete representation of target users.”Persona 是真实用户的虚拟代表,是建立在一系列真实数据(Marketing data,Usability data)之上的目标用户模型。

用户画像，即用户信息**标签化**

什么是用户画像

目录

为什么需要用户画像

互联网的本质是连接

- 人与人
- 人与物
- 物与物

如何在海量的人与物连接起来？

- 理解人
- 理解物

如何理解人？

- 用户画像

为什么需要用户画像

微观

推荐引擎	用于召回、冷启动、兴趣探索等推荐策略
算法模型	作为用户特征，提升推荐ctr、广告ctr等模型的指标
广告引擎	用于广告系统个性化策略等策略
人群包	生成用户人群包，广告主进行划人群投放
运营投放	提供给运营侧精准人群画像，提升运营投放效率

为什么需要用户画像

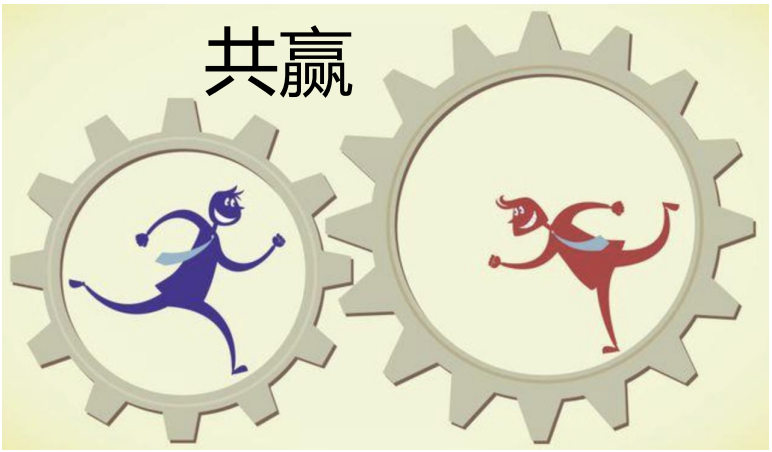
宏观

用户

获取所求
用户体验

企业

用户流量
商业价值



目录

如何构建用户画像


个人信息
年龄
性别
出生日期
学历
职业
手机号
...
用户行为
最后一次使用平台
最近下单时间
最近访问时间
首次访问时间
页面信息
...
用户偏好
商品种类
商品品牌
价格区间
...
用户地区
所在国家
所在省份
所在城市
...
渠道信息
来源域名
来源平台
来源类型
...
设备信息
浏览器
操作系统
IDFV
IDFA
MAC
IMEI
...

如何构建用户画像

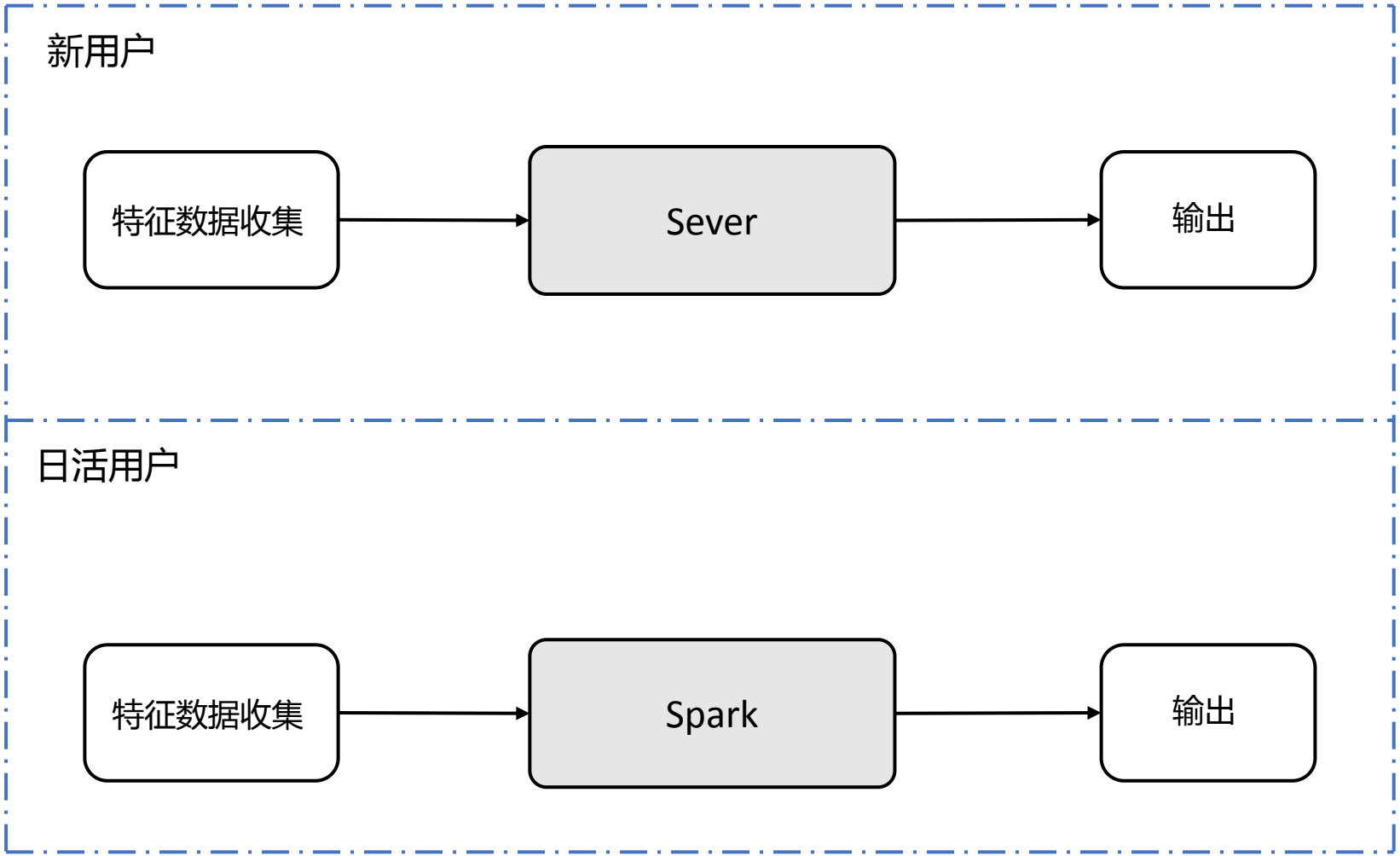
数据层	用户基数数据	用户行为数据
算法层	简单信息提取 （注册时间、渠道等）	规则提取 （金币用户等）
	数学统计 （活跃天数、时长等）	机器学习模型 （年龄、性别等）
输出层	sever	hive

如何构建用户画像

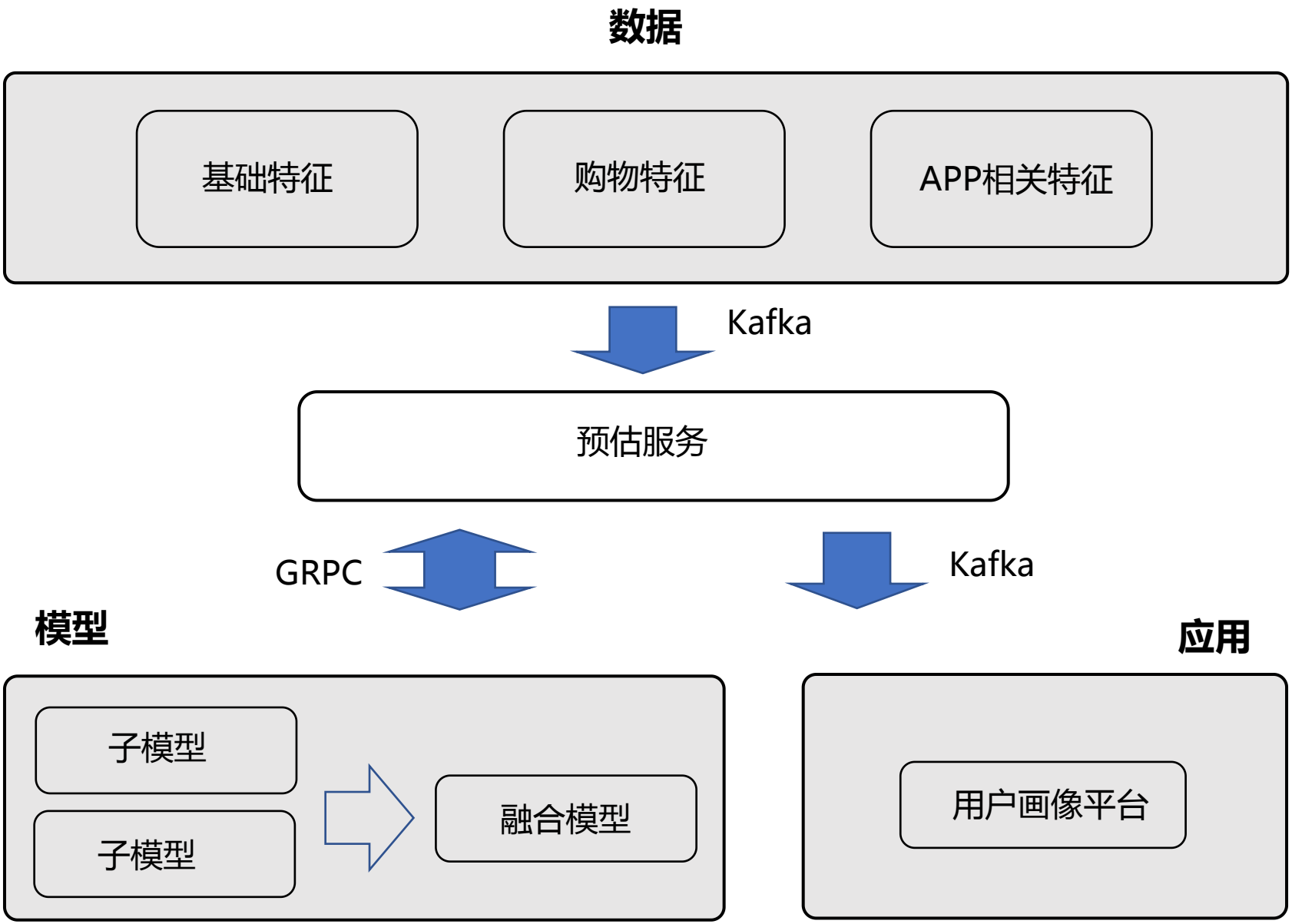
年龄/性别预测

兴趣偏好画像

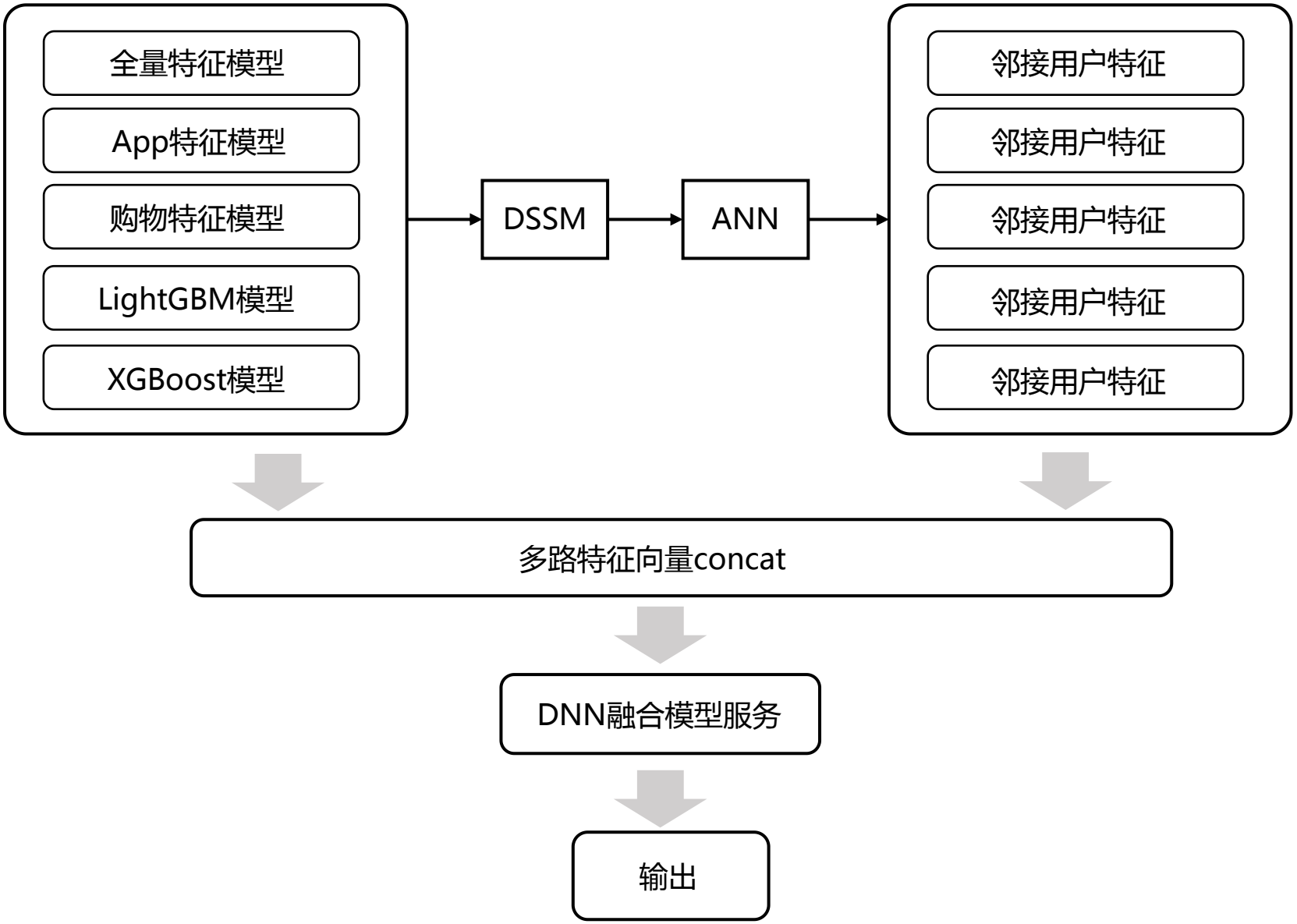
如何构建用户画像-年龄/性别预测



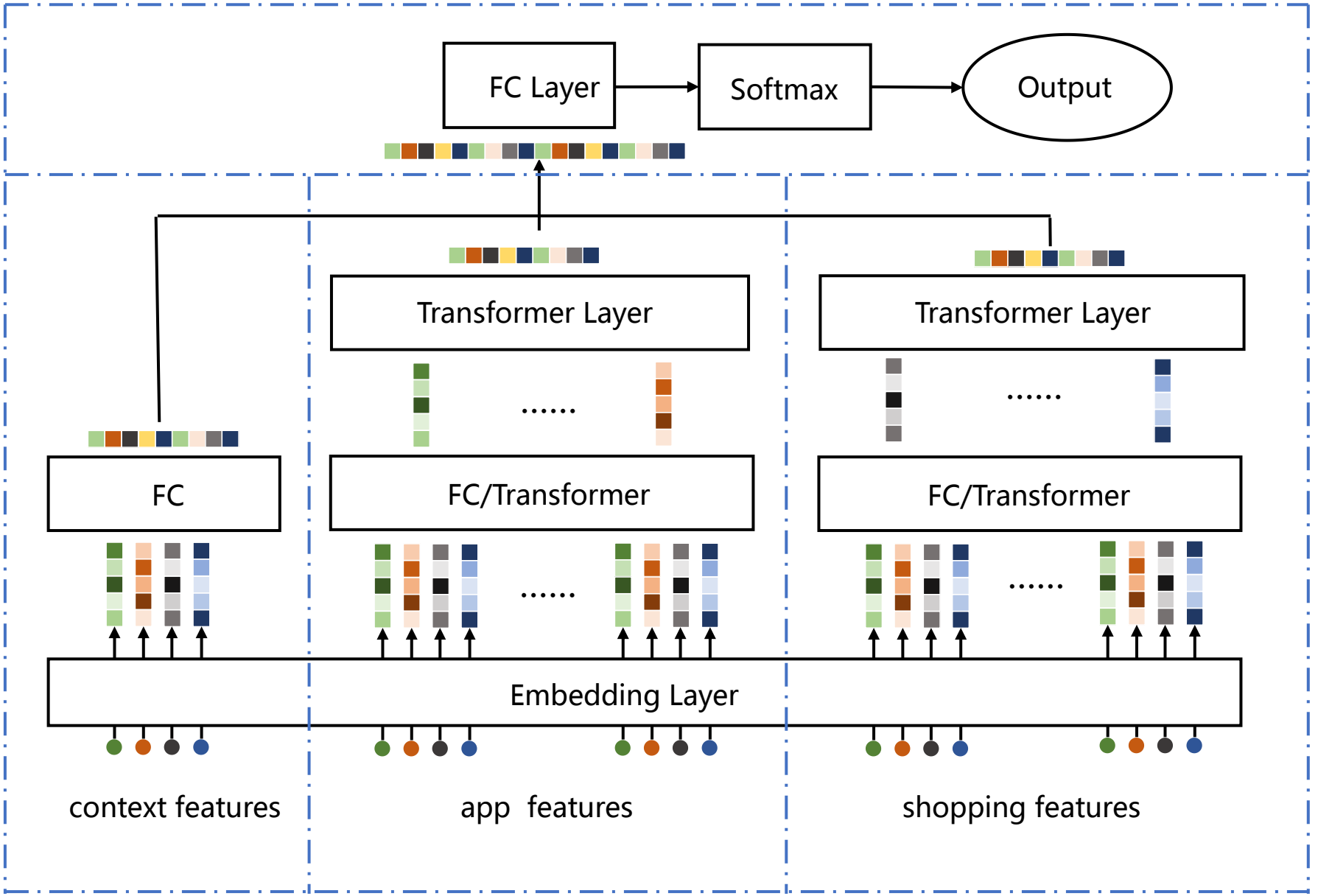
如何构建用户画像-年龄/性别预测



如何构建用户画像-年龄/性别预测



如何构建用户画像-年龄/性别预测



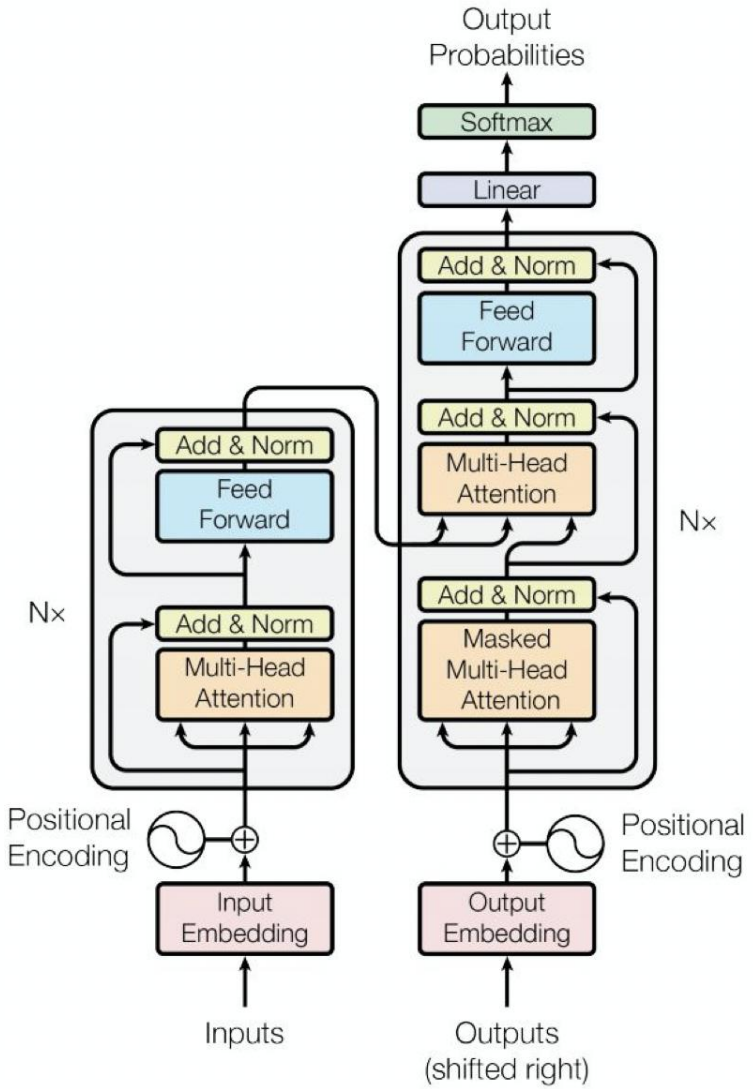
Transformer改动

Multi-Head Attention

Add & Norm

Feed Forward

ADD & Norm



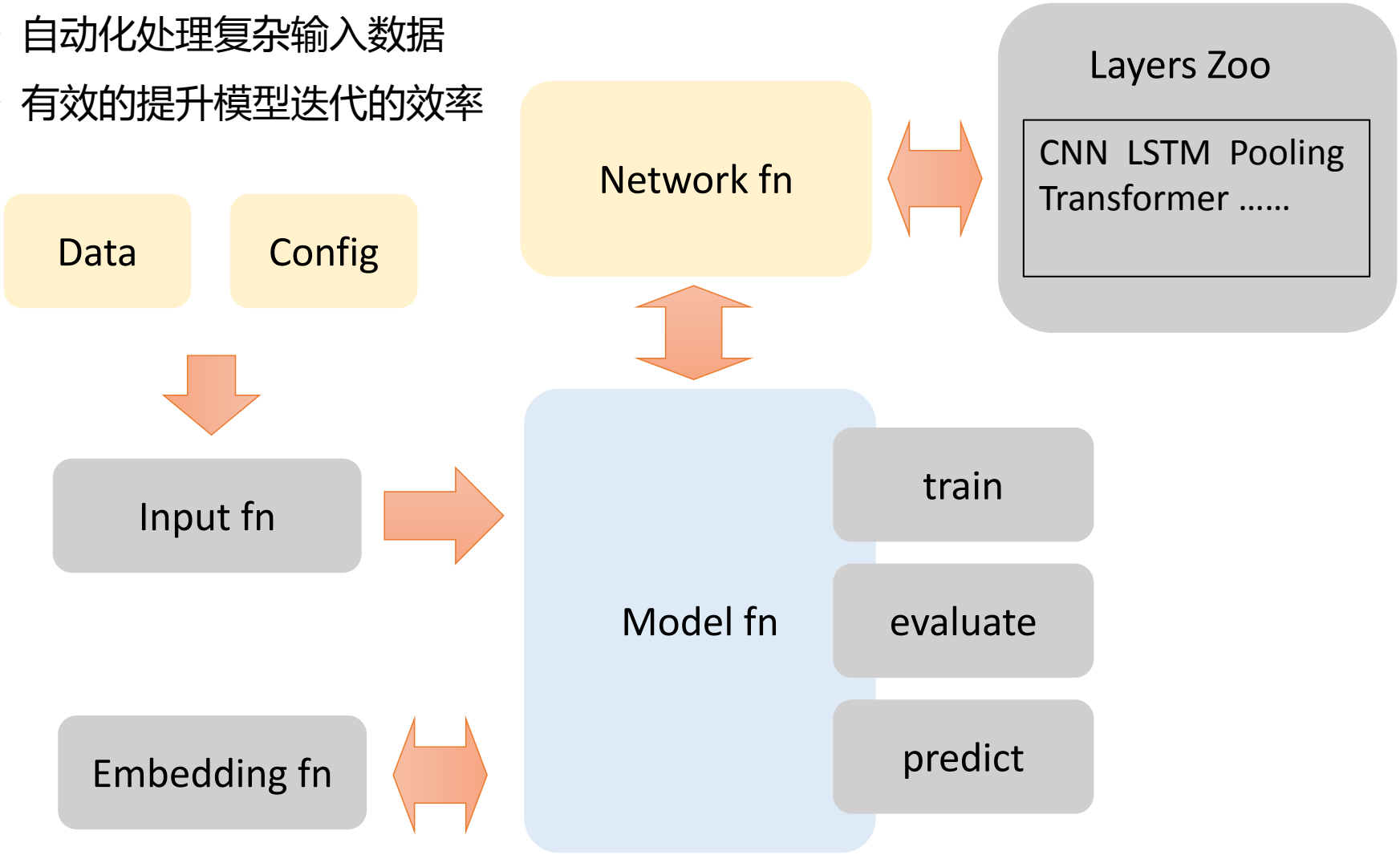
通过实验，只用了最核心的Multi-head Attention

如何构建用户画像-年龄/性别预测

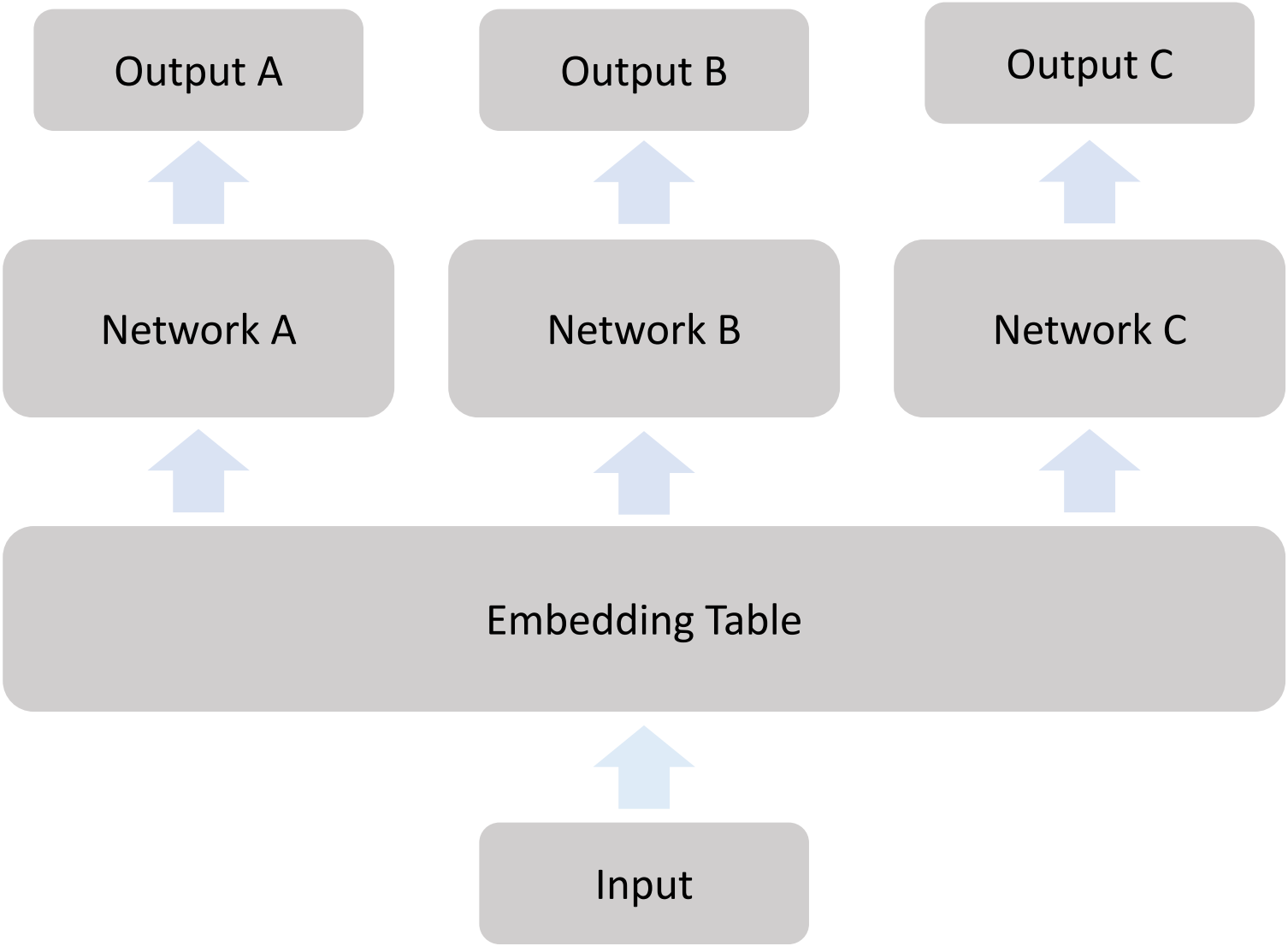
项目设计特征比较复杂，如何提升迭代效率？

沉淀出一套基于TensorFlow Estimator的模型框架：

- 自动化处理复杂输入数据
- 有效的提升模型迭代的效率



多任务学习、迁移学习


Output A
Output B
Output C
Network A
Network B
Network C
Embedding Table
Input

如何构建用户画像-用户偏好画像

时隔7年再次问鼎！马竞加冕队史第11次西甲冠军

原创 · 05/23 02:01 直播吧/新闻频道　　查看原文

直播吧5月23日讯 西甲末轮，马竞2-1战胜巴拉多利德，夺得本赛季西甲联赛冠军！

这是马德里竞技时隔7年再度夺得西甲冠军荣誉，同时这也是马竞队史第11次西甲冠军。

马竞上一次获得西甲冠军，是在2013-14赛季，当时马竞在西甲最后一轮客场1-1战平巴萨，夺得当赛季的西甲冠军。

一级分类：体育

二级分类：足球

三级分类：西甲

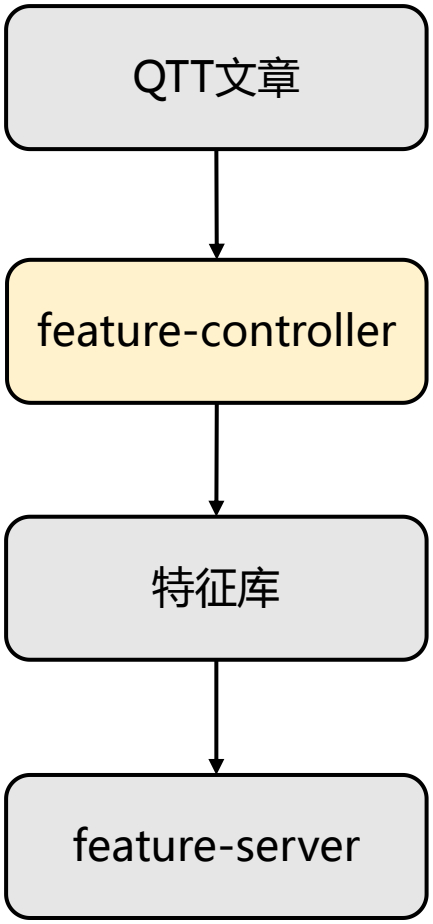
实体词：马竞、西甲

关键词：马竞、西甲、冠军

热点：马竞夺得西甲冠军

……

如何构建用户画像-用户偏好画像



如何构建用户画像-用户偏好画像

一阶特征

基础特征

作者 | 体裁 | 发布时间 | 文章来源

文本特征

实体词 | 关键词 | topic

n-gram | 热点事件 | 文本向量

影视剧名 | 名词提取 | 地域识别

图片特征

人脸识别 | 图片向量

视频特征

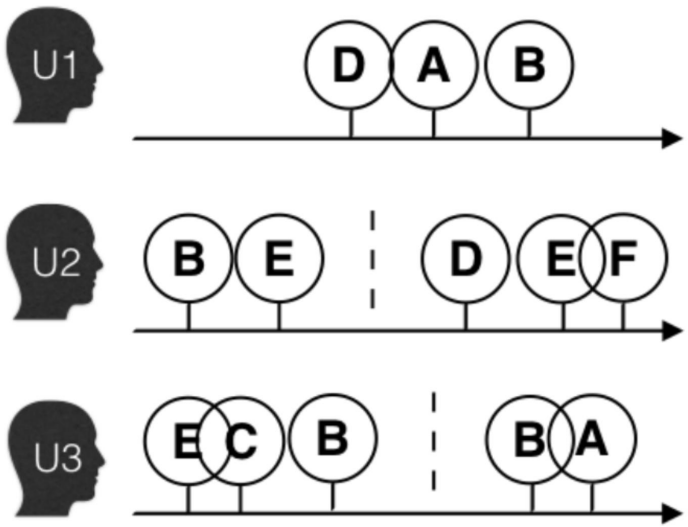
视频向量 | 影视剧名

二阶特征

一级分类 | 二级分类 | 标题党 | 内容质量 | 文章风格

如何构建用户画像-用户偏好画像

在推荐系统中，数量最为庞大的要数偏好类的标签了。平台有多少个物品标签，就会产生多少偏好标签。另一方面，偏好类的标签的产生，依赖于物品标签。因为用户对物品的偏好程度，是通过他对平台物品的曝光，点击，购买等行为计算出来的。



一级类目偏好：体育、影视

二级类目偏好：足球、篮球

三级类目偏好：西甲、湖人

实体词偏好：刘德华、周杰伦

体裁偏好：视频

关键词偏好

topic偏好

内容质量偏好

画像生成逻辑

- 单天画像

$$score_i = w_{action} * w_{tag} * w_{special} * c_{action}$$

- 画像合并

$$score = w_{decay} * score + score_i$$

w_{action} 是用户行为的权重，点赞分享的权重要大一些

w_{tag} tag对应文章的重要度，比如关键词，则是这个关键词关于这篇文章的权重

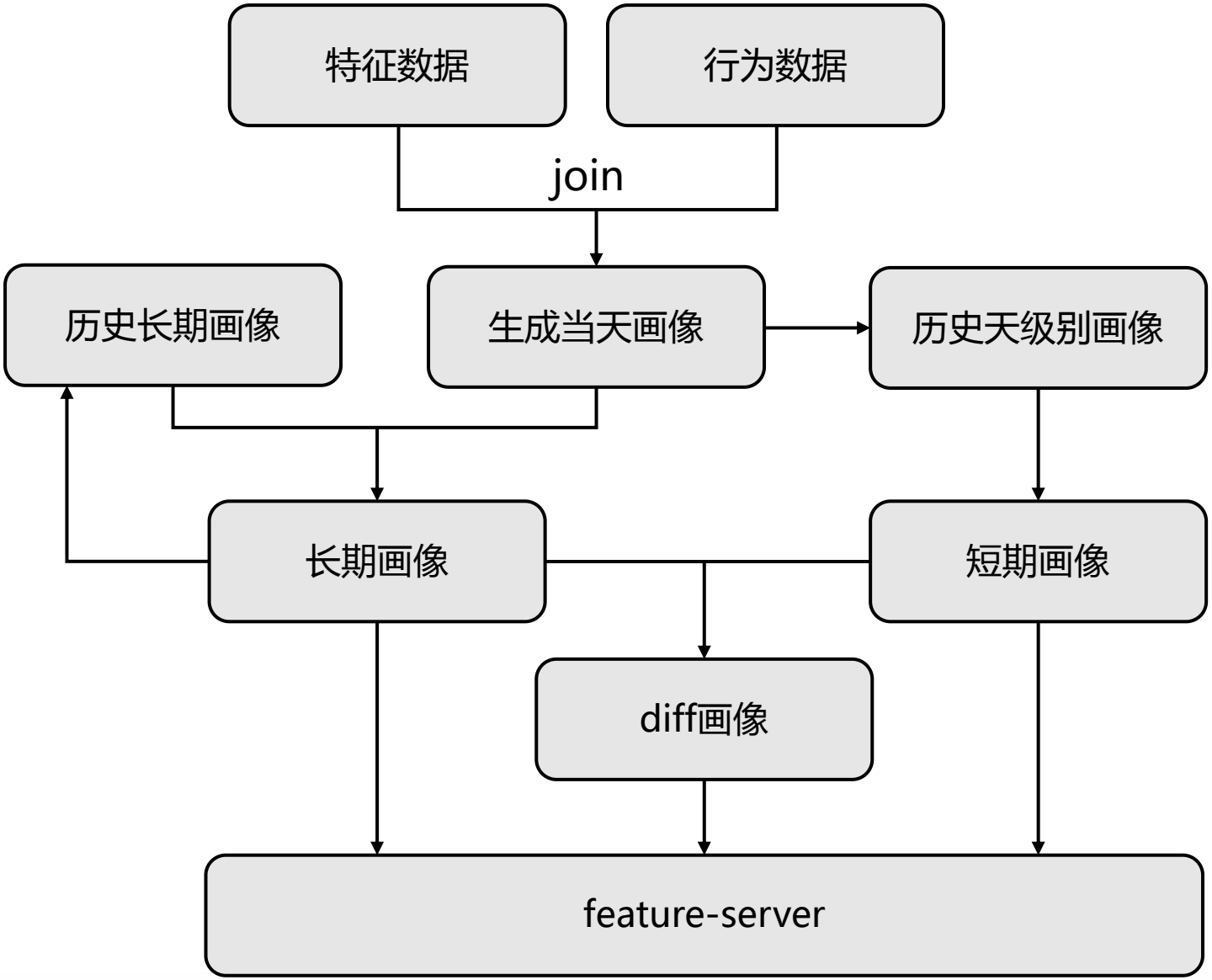
$w_{special}$ 是一个备用的特定调权，针对某些特定的画像进行权重调节，也可以针对某些from的文章的权重进行调节，比如热点

c_{action} 行为的次数

w_{decay} 时间衰减系数。用户的行为会随着时间的过去，历史行为和当前的相关性不断减弱，在建立与实践衰减相关的函数时，我们可套用牛顿冷却定律数学模型。

如何构建用户画像-用户偏好画像

偏好画像构建流程（spark）



如何构建用户画像-用户偏好画像

时间维度

- 长期画像：90天起
- 短期画像：30天画像、7天画像
- 实时画像：session级别
- diff画像：长期画像-30天画像、30天画像-7天画像、长期画像-7天画像

体裁维度

- 图文
- 视频
- 小视频

不同样式

- 单图
- 三图
- 无图

计算逻辑

分类	表达式
点击频次	$pre = click_i$
点击率	$pre = \frac{click_i + m}{show_i + \mathrm{n}}$
正样本分布比例	$pre = \frac{click_i}{click}$
交叉后的相对偏好	$pre^* = \frac{pre}{E(pre)}$
用IDF来体现稀缺度（TFIDF）	$pre^* = pre * log(\frac{\sum_{i=1}^{c} user_i}{user_j + 1})$

目录

目录

总结

推荐模型

700+

用户特征

300+

总结

- 用户画像是我们理解用户的重要手段与方法，只有理解了用户，才能提供更好的服务。
- 对于推荐系统而言，偏好画像是重点，数量上占了推荐系统用户画像的绝大多数，是我们召回和模型训练的基石。
- 用户embedding也是用户画像中重要的一种用户标签

麦思博(msup)有限公司是一家面向技术型企业的培训咨询机构，携手2000余位中外客座导师，服务于技术团队的能力提升、软件工程效能和产品创新迭代，超过3000余家企业续约学习，是科技领域占有率第1的客座导师品牌，msup以整合全球领先经验实践为己任，为中国产业快速发展提供智库。

高可用架构公众号主要关注互联网架构及高可用、可扩展及高性能领域的知识传播。订阅用户覆盖主流互联网及软件领域系统架构技术从业人员。

高可用架构系列社群是一个社区组织，其精神是“分享+交流”，提倡社区的人人参与，同时从社区获得高质量的内容。