

主要内容

- 知乎推荐页场景和 Ranking 历程介绍
- 深度学习在 Ranking 中的尝试和应用现状
- Ranking 面临问题和未来研究方向

知乎推荐页场景



为什么电动汽车电池不能精确显示剩余电量?

赵玮炜: 首先回答楼主的问题"为什么 不能像手机一样显示剩余电量百分比, 这个问题的答案是: 能显示百分比, ...



17 特同

靠工资在北京买房现实吗?

羊迪: 当然现实了, 不需要年入百万, 月入一万都没多大问题,来,我教你。 首先,确认下购房条件是否满足,如...



388 赞同·136 评论

后悔当初没有考上高校?环球网校轻松考取MBA, 点击报名

名校梦想无法实现? 现在机会来了! 环球网校轻松考取 MBA, 帮你实现名校学历, 职场转型利器, 薪资爆炸式提升

环球网校 的广告· 查看详情

据说机器学习长于预测, 计量经济学长于解释。有什 么具体例子是机器学习完成了很好的预测却在解...

颢卿: 机器学习之于计量经济学 - 随机森林算法大战最 小二乘法 今天这篇正好借花献佛,来和大家从解释和预 测两个角度来和大家简单讲讲机器学习和传统计量方法... 242 赞同 · 21 评论

1.783 赞同·90 评论

23:19 HD (\$ 490)

Q 熊晓鸽的投资方法

推荐

0 赞同

执榜

▲ 提问 视频

Deeplearning.ai卷积神经网络(1.6 如何做三维卷 积)

忆臻: 0、前言 打算对吴恩达老师的 Deeplearning课程做一个笔记回顾, 大体整理思路为吴老师课堂内容和个...



提升学历很难吗? 来环球网校考在职研究生, 学时 短,含金量高

学历是求职的敲门砖, 也是升职加薪的垫脚石。高考失利没 关系! 在职研是你的又一次机会。来环球网校, 帮你高效学 习,提升学历!

环球网校 的广告·查看详情

电动车续航破600,燃料电池汽车出路何在?

玖牛咨询: 导读 电动汽车与燃料电池汽车谁是新能源汽 车未来? 比亚迪唐EV、现代NEXO、丰田MIRAR续航里程 接近。失去续航里程高的优势, 氢燃料电池汽车出路何...

1 赞同·20 评论

海洋里有哪些有趣的哺乳动物?

一个男人在流浪: 谢激。因为题目里说 到了是海洋哺乳动物,这就把我们的可 选范围限制在了一定很小的面里--...



23:55

毕赣邀你下周日见面聊爱情

推荐

热榜

视频

(45)

▶ 提问

怎样评价一个演员的台词功底?



大酱: 我当时以为这是真的杀人犯采访

2.757 赞同 · 447 评论

有什么事是外国人来了中国之后才知道的?

笑笑木: 很多国家是不"拼假"的。比如, 今年国庆假期之 前的周六日正常上班上学, 然后连着放七天, 这就是"拼 假"。这种模式现在被很多人吐槽。我上大学的时候,隔...

216 赞同 · 77 评论

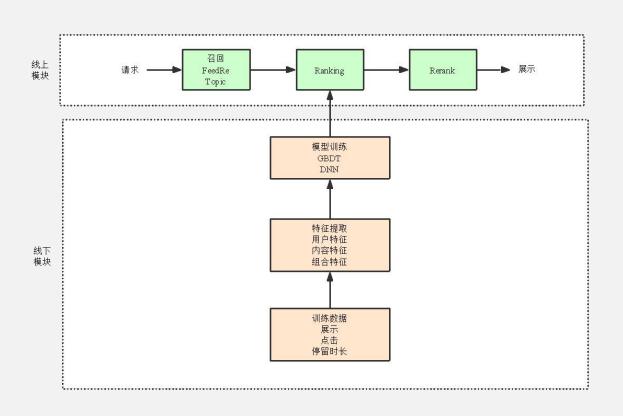
汽车的操控性到底是指什么?

翰德林曼: 汽车的操控性到底是指什么? 说到操控性, 在 中国和欧美还是有些差异,在中国,普通消费者理解的操 控包含了动力加速性能,驾驶性,转向,极限操控,甚至...

25 赞同

如何克服一个人独处时草名的空虚寂寞感?或许, Soul里有要的答案

推荐页请求流程



推荐页模块详解

- 召回:负责把用户可能感兴趣的内容提取出来,重点是全
 - ▶ 基于话题: 关注, 行为挖掘
 - ▶ 基于内容: 协同
- 排序:负责对召回的内容进行打分,可以理解为感兴趣程度,重点是准
 - ▶ 基于规则:时间顺序,线性加权
 - ▶ 基于模型: GBDT, DNN
- 重排序:出于产品或者业务的考虑,对排序后的内容进行重排,最终展示
 - ▶ 提权:比如对视频进行一定的提权
 - ▶ 隔离:相似内容隔开
 - ▶ 强插: 高质量的新内容流通

推荐页Ranking历程



模型选择



特征介绍

- 用户画像
 - 用户属性特征: 性别等
 - 》 统计特征,用户点赞数等
- 内容画像
 - 固有特征:文章长度,关键词等
 - > 统计特征: 历史点赞数等
- 交叉特征
- 用户与内容的交叉特征: 比如用用户感兴趣的话题和当前待推荐内容的话题交叉

特征介绍

●特征类别

- > 数值特征:文章长度,点赞数
- > Onehot: 内容类型
- > Multihot: 内容多个话题 id
- > Onehot with value: 用户对单类型内容的感兴趣程度
- > Multihot with value: 用户对各话题的感兴趣程度

特征设计

●设计原则

- ▶ 特征尽量全:从现有的数据中提取尽可能多的特征
- ▶ 特征原始值全: 比如加历史CTR 特征的时候, 可以把 pv 和 click 都带上
- ▶ 覆盖率大:去掉一些覆盖率很低的特征,这些特征影响影响范围小,大部分是缺失值
- » 线上线下一致:覆盖率和取值分布尽可能接近

●新特征方向

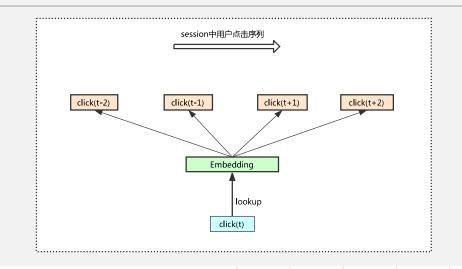
- → 显式交叉特征: DNN 能学习特征的非线性能力,增加交叉特征可以降低模型搜索的空间,在训练数据一定的情况下可以提升效果,如用户的话题兴趣和当前话题的均值和最大值,效果提升明显
- » 出于业务考虑:需要对业务有一定的理解,把自己当做用户,考虑什么情况下点击率会大,什么样的内容更容易被用户点,比如视频在 wifi 下更容易被点,视频点击率高的人更喜欢视频
- » 数据挖掘特征:如内容 Embedding 特征

特征设计

- ●内容 Embedding 介绍
 - ▶ Embedding 目的: 把内容映射到低纬空间,相似内容距离较近,可以当做内容特征
 - > 文本角度: tfidf, 关键词进行word2vec 等
 - » 行为角度: 考虑用户在知乎的行为,搜索内容相关性较好,依据搜索点击序列设计 Embedding
- 实现细节
 - > 数据:将搜索行为进行session切分,组织成类似于 sentence 序列
 - » 样本: 85亿
 - ▶ 模型: skip-gram
 - > loss: nce-loss

特征设计

• skip-gram 结构



embedding 效果

ID

凹合	97421005	四合: 1	家VD VC Ve	这件的涨键	<u>品, </u>	大大吃吗?	谢趣,	这些保健品	AB连非是	官谷易畎之	的,可以称	允。维生	系し和维生	系L小人谷	汤	多吃疏	米刀
回答	67962958	回答:	开始拼命地	吃维生素b	族 vb3 vb	o6 vc ve 望能治疗	答主:	怎样私信你	? 我也遇	到同样问	题,痘痘长	了大概有一	一年了,	今年25周	岁, <mark>期</mark> 间	试过各种	方
回答	61376154	回答:	我是22岁的	女生,请问]日常吃作	十么营养品比较如	我高三	的时候被我	妈逼迫吃	葡萄籽、	/c和各种保	健品,但是	是葡萄籽我 :	最讨厌吃,	因为真	的很难吃	, i
回答	2927800	回答: 组	继续用综合	维生素还是	只服用v	c vb?	题主你	好! 综合维	生素的问]题之前我	也一直在关	注,首先陈	东述我的观点	点1.综合组	生素按	时服用后	, 🧦
回答	66482293	回答:[吃维生素bb	疾 vb3 vb6	/c ve 望能	能治疗我的痘痘:	我也吃	过小瓶的VE	36坚持吃	了三个月数	好像 基本上	就是按说	明上的吃的	吃了10	佤 吃完品	后确实能!	控制
回答	93548384	回答:	看微博里说	纽斯特的V	C+VE+奮	葡萄籽可以美白,	本人黑	,双11买的	纽斯特的	美白套餐	,就是题主	说的那个。	到现在四·	个月了。i	首先说白	,没感觉	变
回答	83179992	回答:「	吃药店的小	瓶vc和吃自	然堂, 注	<u> </u>	没有差	别,因为你	摄入的维	生素剂量	是一样的,	这是标准剂	引量,只是1	自然堂、清	 	等保健品	品
回答	94153709	回答: 7	想利用服用	VC和VE的	胶囊美白	1, 这科学吗? V	两种维	生素都有抗	氧化的功	能,至于	能不能美白	,我只能说	的有待商榷。	维生素A	DEK属	F脂溶性组	维生
回答	18055293	回答:[因为长痘痘	,可以VB.	VC.VE-	起服用一段时间	1 我当时	也是长过很	多痘痘,	在杭州看	的中医调理	一段时间研	角实好了,「	乞中药的阿	司时医生	配了vc v	b纺
回答	93241351	回答:「	吃富含vc的	水果和直接	喝vc片有	有什么区别?	谢邀!	当然会有区	别! 片剂	有提取的	和合成的两	种片剂含量	量多少不一人	十剂中营	卡素单一	水果中富	含:
回答	33467535	回答:	普通人长期	每天吃 Na	ure Mad	le莱萃美的vc和v	€ 谢邀,	首先抱歉的	回答,因	为对您购	买的保健品	并不了解,	无法正面[回答您的问	可题,抱	歉。其次	ζ,

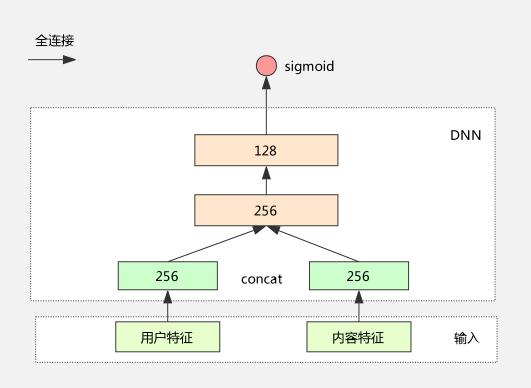
CTR 模型

●选择 CTR 模型原因:

- ▶ 推荐页排序目标是把用户推荐感兴趣的内容排在前面,可有下面两个学习目标
 - ◆ 停留时长:适合用回归问题来解决,最后会偏向于长文章
 - ◆ 点击率:二分类问题,知乎的问答一般不长,更加合适
- 分类问题相比回归问题,目标类别少,相对准确率高
- > 分类问题场景业界应用较广,可交流空间大
- > 分类问题最后会输出一个概率分,方便与多目标结合
- ●损失函数

$$Loss = -\sum_{i=1}^{N} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

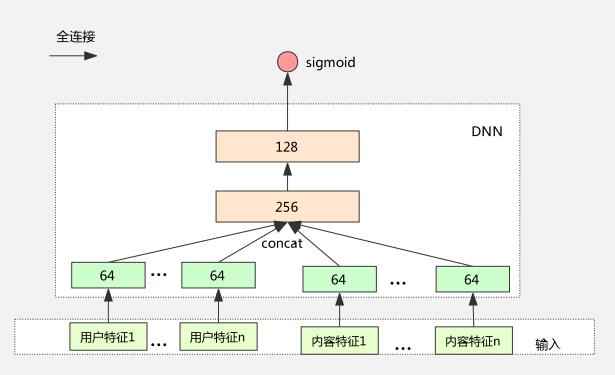
最初DNN 结构



- 1. 将输入特征分为用户和内容两块
- 2. 经过特征映射后分别通过全连接 与两个独立的隐含层连接
- 3. 两个独立的隐含层 cancat 后再 经过两个全连接层
- 4. 最后输出 sigmoid 与交叉熵损失 作为 loss

AUC: 0.7618

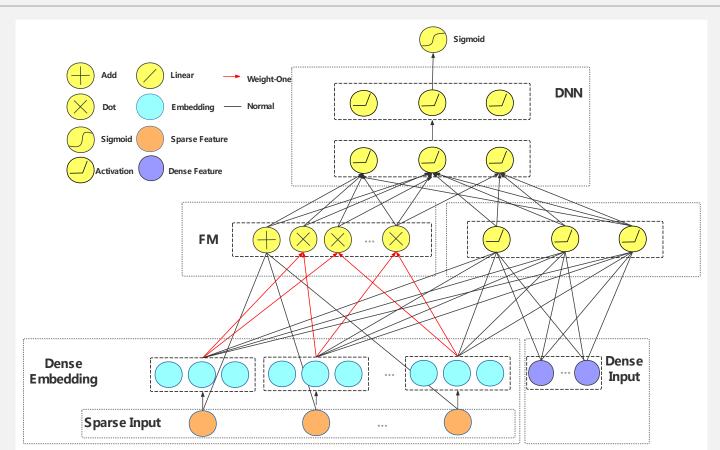
优化 DNN 结构



- 1. 将用户和内容的特征,分别按照 内容的 field 分为不同的 block
- 2. 每个 block 先经过全连接到独立 的隐含层
- 3. 将上面的隐含层 concat 再经过 后面的 DNN 模型

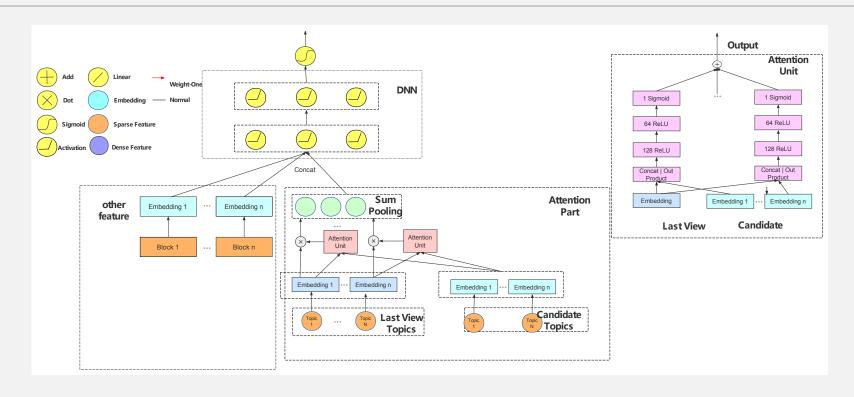
AUC: 0.7678, 提升0.6%

Deep FM



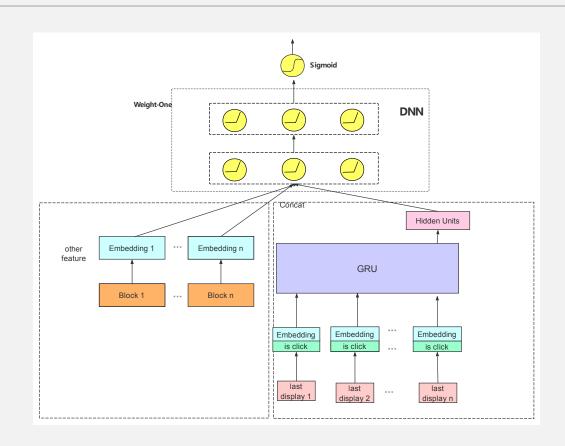
- 1. 增加了一阶和 FM 模 块,FM 通过 block 之间的内积实现
- 2. AUC 提升 0.2%

Last View + DIN



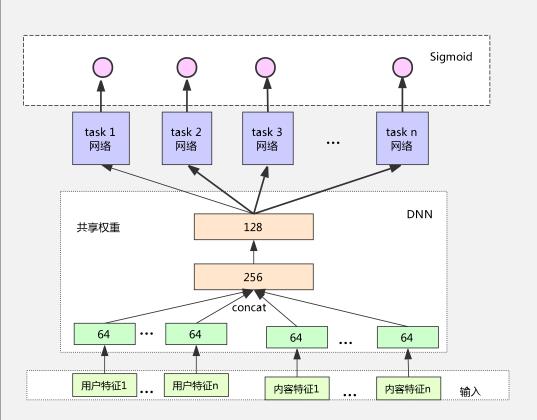
- 1. Last view topic 与当前内容的几个 topic 计算 Attention Score, 再按权重进行 sum pooling
- 2. AUC 提升约 0.2%

Last Display + GRU



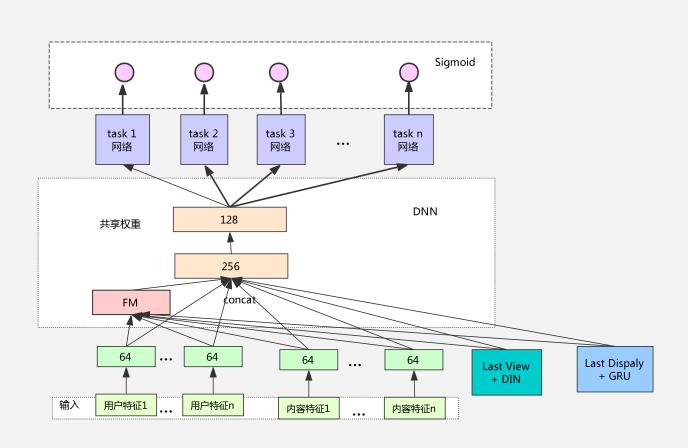
- 1. Last Display 经过 Embedding 后与是否点击结合,再进入 GRU 模块,最后状态当做 DNN 输入
- 2. AUC 提升约 0.4%

多目标



- 1. 每个 task 共享前面的几层权重,可以 节省训练和预测的计算量
- 2. Loss 可以是几个 task 的 loss 做简 单线性加权
- 3. 上线仍然要确定各个 ctr 的加权值, 经验比较重要
- 4. 上线后线上表现:点击率基本不变,而 其他的几个指标,比如点赞,收藏大幅 提升

最终模型结构



经验分享

- 对于随时间变化的统计特征,比如用户和内容画像的统计值,线上 service 应当纪录请求时的值,生成训练样本时直接从纪录的日志里获取,避免特征穿越问题;
- 如果发现线下效果好,比如 AUC 和 NDCG 提升明显,但上线效果不显著,很可能是特征不一致导致的,可重点排查;
 - 线上线下最好使用同一套特征抽取框架,只需使用的相同特征配置便可保证一致性,我们 Global Ranking 使用同一套 proto 结构和特征抽取模块实现;
- 做特征归一化操作,发现有特别大的值,比如几万或者几十万,要先取 log ,不然会导致这个特征大部分值都 趋向0,相当于特征失效;
- 输入特征要做非法检查,防止出现 inf, nan, 而导致模型训练出现异常的参数; ● 对于线上的每次请求, 用户特征都是一样的, 可以只计算一遍用户特征相关的 block, 避免冗余运算;
- 训练数据量要尽可能大,可以使用 FlatBuffer 结构把训练数据存放在 HDFS 上,训练时直接从 HDFS 读取, 边读取边训练;
- 线上模型要能自动更新,过老的模型效果下降严重;

面临问题

- 推荐页与搜索页的特性不同
 - ▶ 搜索带着 query 来的,结果与之相关性越高越好,不用太关心结果的多样性
 - ▶ 推荐页用户没有明确的目的,但是有兴趣偏好和对结果的多样性需求,推荐既要准确又要多样化
- CTR 预估模型是 pointwise 模型,没有考虑单个内容与其他内容同时出现的影响
- 用户对感兴趣的东西会出现审美疲劳,要及时抓住这种特点,比如一个算法工程师看完几个机器学习文章后就不想再看了,这时候要能推荐一些其他话题的内容

未来方向

- 强化学习
 - Actor: 根据用户过去的浏览和点击行为生成推荐页整屏结果
 - Critic: 接收到点击或者其他正向行为作为 reward, 同时训练 Critic 和 Actor 网络参数
- 优点
 - ▶ 能及时捕捉用户的反馈,从而避免对同一话题产生审美疲劳
 - ▶ 推荐整屏幕内容,避免 pointwise 方式下内容较为集中问题
- 缺点
 - 模型结构复杂,模型参数训练较困难

关注我们



智能推荐系统

