


 所谓 (/users/845813) 2020-10-30 18:02:21 (最初创作于: 2020-10-30 18:00:07) 17 阅读 1 收藏 (/articles/185347/mark/)

知识体系: [数据挖掘 \(/articles/?kid=944\)](#) [搜索推荐 \(/articles/?kid=945\)](#) [人工智能 \(/articles/?kid=951\)](#) [数据处理 \(/articles/?kid=955\)](#) [数据建模 \(/articles/?kid=957\)](#)  修改知识体系

文章标签: [算法 \(/search?q=算法&type=ARTICLE\)](#)  修改标签  标签历史 (/articles/185347/tags/history)

附加属性: [作者原创](#)

搜索深度精排大模型优化资料汇总

搜索深度精排大模型优化资料汇总

我们将搜索/推荐精排大模型的优化分为四大模块,

1. Loss设计;

把Loss的设计排在特征工程等操作之前,结合业务指标并设计对应的Loss是模型优化的第一步, 如果Loss设计不佳,可能后续的诸多结论在新的情况下都会有问题, 所以先尝试重点优化这一模块;

2. 数据预处理;

2.1. 数据清洗:在我们的数据中,存在非常多的脏数据,这些数据的处理可以帮助我们更好地提效;

2.2. 数据采样:因为大模型这块数据量是巨大的,所以需要有合理的采样策略;

3. 特征工程;

特征工程目前还是建模中最为核心的一块,也是提效最快简单的部分;

4. 模型结构;

4.1.交叉特征优化, 特征的交叉是深度模型的核心之一, 可以用作横向特征信息的捕捉;

4.2.序列信息挖掘, 序列特征的挖掘, 可以用作纵向信息的捕捉;

4.3.细节优化,包括模型单元先后调用等.

5. 其他探索

论文调研,杂谈,有趣的论文,隐式信息探索等等。

Loss设计

目前内网关于深度大模型的设计大多是基于DeepMTL的模式,关于这种模式我们进行了对应的调研,对应的调研内容如下:

因为我们的评估指标是从曝光到购买的转化率,所以我们线下的优化指标为PVP的GAUC,基于调研我们最终设计了目前相对较优的方案,参见下面的文章[14]

个人ATA学习汇总:

- 1. 多任务学习1: 为什么要进行多任务学习? (ArXiv17) (https://www.atatech.org/articles/183907)
- 2. 多任务学习2:NMTR_Neural Multi-task Recommendation from Multi-behavior Data (ICDE19) (https://www.atatech.org/articles/184397)
- 3. 多任务学习3:RecSys20 Best Paper, PLE, Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations (https://www.atatech.org/articles/184396)
- 4. 多任务学习4:Improving One-class Recommendation with Multi-tasking on Various Preference Intensities(RecSys20) (https://www.atatech.org/articles/184419)
- 5. 多任务学习5: MTL实践经验调研(内容来自于国外Blog的分享) (https://www.atatech.org/users/845813/own/?p=1)

6. 多任务学习6:Modeling Consumer Buying Decision for Recommendation Based on Multi-Task Deep Learning(CIKM18) (<https://www.atatech.org/articles/184576>)
7. 多任务学习7: SEMAX_Multi-Task Learning for Improving Recommendations(IEEE19) (<https://www.atatech.org/articles/184577>)
8. 多任务学习8:Improving Entity Recommendation with Search Log and Multi-Task Learning(IJCAI18) (<https://www.atatech.org/articles/184647>)
9. 多任务学习9:MKR_Multi-Task Feature Learning for Knowledge Graph Enhanced Recommendation(WWW19) (<https://www.atatech.org/articles/184648>)
10. 多任务学习10:DBMTL_Deep Bayesian Multi-Target Learning for Recommender Systems(Arxiv19) (<https://www.atatech.org/articles/184726>)
11. 多任务学习11:Jointly Learning Explainable Rules for Recommendation with Knowledge Graph(WWW19) (<https://www.atatech.org/articles/184911>)
12. 多任务学习12:Why I like it_Multi-task Learning for Recommendation and Explanation(RecSys2018) (<https://www.atatech.org/articles/184912>)
13. 多任务学习13: 多任务学习(MTL)在搜索推荐场景的应用调研&实践尝试&小结 (<https://www.atatech.org/articles/184917>)
14. 多任务学习14: 多任务学习(MTL)在搜索推荐场景的调研&实践&扩展 (<https://www.atatech.org/articles/185228>)



集团与CTR/CVR大模型相关的ATA好文:

1. GMSL: GRU-based Multi-Stage Learning for E-commerce Ranking 电商排序多阶段学习模型 (<https://www.atatech.org/articles/150857?spm=ata.13269325.0.0.66a649faCwLYbA>)
2. 1688搜索转化优化之CTR/CVR模型 (<https://www.atatech.org/articles/173972?spm=ata.13269325.0.0.66a649faCwLYbA>)
3. Two-Step Relation Network: 飞猪个性化在CTR/CVR联合预估上的探索 (<https://www.atatech.org/articles/172923?spm=ata.13269325.0.0.66a649faCwLYbA>)
4. ICBU Galaxy1.0: 基于ESMM的CTR&CVR联合训练模型实践 (<https://www.atatech.org/articles/179335?spm=ata.13269325.0.0.270149faZrLBjw>)
5. AE搜索Deep CTR/CVR探索: 稀疏样本利用和特征感知建模 (<https://www.atatech.org/articles/155157?spm=ata.13269325.0.0.111949fa5E5aAt>)
6. 搜索算法国际化之马六甲项目 - CTR/CVR 大模型篇 (<https://www.atatech.org/articles/125697?spm=ata.13269325.0.0.1b6549fa0DJAsv>)
7. Deep Bayesian Multi-Target Learning: 淘宝直播推荐rank模型上的创新 (<https://www.atatech.org/articles/134855?spm=ata.13269325.0.0.62d249fadlJig>)
8. 大规模精准ctr/cvr预测模型: tmall搜索双11总结 (<https://www.atatech.org/articles/124033?spm=ata.13269325.0.0.1b6549fa0DJAsv>)
9. ICBU搜索排序多任务多目标优化 (<https://www.atatech.org/articles/164481?spm=ata.13269325.0.0.1b6549fa0DJAsv>)
10. 大药房首页feed流多目标学习实践 (<https://www.atatech.org/articles/176399?spm=ata.13269325.0.0.172749faCD5TIS>)

数据预处理

数据清洗

个人ATA学习汇总:

- 线下实验中,暂未总结

集团与数据质量相关的ATA好文:

1. ICBU Galaxy2.x: 基于实曝光数据和模型优化的排序模型实践 (<https://www.atatech.org/articles/179336?spm=ata.13269325.0.0.6d3b49fa2bFyPq>)

数据采样

个人ATA学习汇总:

- Pending

集团与数据质量相关的ATA好文:

- Pending

特征工程

特征工程目前还是现在模型的灵魂之一,是最简单提效的方案。

个人ATA学习汇总:

1. 搜索推荐相关竞赛的方案总结_IJCAI18冠军方案总结 (<https://www.atatech.org/articles/176757>)
2. 搜索推荐相关竞赛的方案总结_IJCAI18亚军方案总结 (<https://www.atatech.org/articles/176642>)
3. ATA搜索精排大模型特征汇总(持续更新ing) (<https://www.atatech.org/articles/176820>)

集团与特征工程相关的ATA好文(与电商搜索相关性大的):

1. 淘宝用户类目偏好算法 (<https://www.atatech.org/articles/52080?spm=ata.13261165.0.0.57c428e2OBrLQb>)
2. UGC优质内容挖掘 (<https://www.atatech.org/articles/1296763>)
3. 搜索算法国际化之马六甲项目 - CTR/CVR 大模型篇 (<https://www.atatech.org/articles/125697?spm=ata.13269325.0.0.4da249fapNIWNw>)
4. GMSL: GRU-based Multi-Stage Learning for E-commerce Ranking 电商排序多阶段学习模型 (<https://www.atatech.org/articles/150857?spm=ata.13261165.0.0.46db652eBZ9kNo>)
5. 飞猪猜你喜欢新特征体系实践 (<https://www.atatech.org/articles/149756>)
6. 旅行周期性推荐实践 (<https://www.atatech.org/articles/147230?spm=ata.13261165.0.0.291b2cd0IKdaTi>)
7. 用户画像在Ctr Model中的应用和挑战 (<https://www.atatech.org/articles/156785>)



模型结构

神经网络的模型结构设计有很多地方是和特征工程相辅相成的,比如基于网络层的交叉和特征工程中的交叉特征有很多类似的地方,不过是从网络层角度除法进行建模, 我们将其暂时划分为三大块。

网络交叉

特征的交叉是深度模型的核心之一, 可以用作横向统计特征信息的捕捉,基于调研我们最终设计了目前相对较优的方案(尚未公布),关于各个模型的比较可以参见文章[10];

个人ATA学习汇总:

1. 特征交叉1: DeepFM实现 (<https://www.atatech.org/articles/177191?spm=ata.13261165.0.0.cf06473534T9bg>)
2. 特征交叉2:Holographic Factorization Machines(AAA119) (<https://www.atatech.org/articles/179599>)
3. 特征交叉3:TFNET:Multi-Semantic Feature Interaction for CTR Prediction(SIGIR20) (<https://www.atatech.org/articles/179836>)
4. 特征交叉4:AutoFIS:Automatic Feature Interaction Selection in Factorization Models for CTR Prediction(KDD 2020) (<https://www.atatech.org/articles/179992>)
5. 特征交叉5:NFM:Neural Factorization Machines for Sparse Predictive Analytics(SIGIR17) (<https://www.atatech.org/articles/180168>)
6. 特征交叉6:ONN:Operation-aware Neural Network for User Response Prediction(Arxiv19) (<https://www.atatech.org/articles/180651>)
7. 特征交叉7:AFM:Attentional Factorization Machines_Learning the Weight of Feature Interactions via Attention Networks(IJCAI17) (<https://www.atatech.org/articles/180743>)
8. 特征交叉8:AoAFFM:Attention-over-Attention Field-Aware Factorization Machine(AAAI20) (<https://www.atatech.org/articles/180772>)
9. 特征交叉9: xDeepFM:Combining Explicit and Implicit Feature Interactions for Recommender Systems(KDD18) (<https://www.atatech.org/articles/180955>)
10. 特征交叉10:CTR神经网络特征交叉汇总&尝试&思考 (<https://www.atatech.org/articles/181082>)
11. 特征交叉11:FuxiCTR_An Open Benchmark for Click-Through Rate Prediction (Arixiv20) (<https://www.atatech.org/articles/183172>)

集团提到特征交叉的ATA好文:

1. 基于Transformer+xDeepFM的核销率预估模型在88VIP购后权益推荐中的应用 (<https://www.atatech.org/articles/168959?spm=ata.13269325.0.0.168e49faSG6klt>)

序列信息挖掘

个人ATA学习:

1. 序列数据挖掘1: HGN:Hierarchical Gating Networks for Sequential Recommendation (KDD19) (<https://www.atatech.org/articles/181371>)
2. 序列数据挖掘2:CosRec_2D CNN for Sequential Recommendation(CIKM19) (<https://www.atatech.org/articles/183912>)

集团提到特征交叉的ATA好文:

- 目前暂未调研

细节调优&模型内存优化等

调参相关

1. 正确的Normalization很重要! ! ! (ArXiv20) (<https://www.atatech.org/articles/179295>)

Embedding内存问题

1. Embedding优化学习1:Compositional Embeddings Using Complementary Partitions for Memory Efficient Recommendation Systems(KDD20) (https://www.atatech.org/articles/185276)

其他探索(杂文笔记)

此处是一些和业务直接优化相关性没那么直接的文章,算一些杂文

一.CVR优化:

1. Survival Analysis知识学习 (https://www.atatech.org/articles/178977)
2. Modelling Delayed Feedback in Display Advertising学习 (https://www.atatech.org/articles/177950)
3. An Attention-based Model for CVR with Delayed Feedback via Post-Click Calibration(IJCAI20)学习 (https://www.atatech.org/articles/178975)

二. 一些有趣的顶会论文

1. On Sampled Metrics for Item Recommendation (KDD20) (https://www.atatech.org/articles/182200)

三. 会议论文汇总

1. 会议论文汇总1:AAAI2020 25篇推荐系统相关论文&摘要 (https://www.atatech.org/articles/177864)
2. 会议论文汇总2:Recsys2020_39篇长文汇总 (https://www.atatech.org/articles/185346)

四. 杂文

1. 搜索引擎&推荐系统 (https://www.atatech.org/articles/176441)
2. 电商搜索引擎的一些思考 (https://www.atatech.org/articles/176694)
3. 搜索推荐业务必备词汇表V1 (https://www.atatech.org/articles/183909)

评论文章 (1)

 3 (/articles/185347/voteup)

 0



他们赞过该文章

合吉 (/users/965862) 银汉 (/users/873599) 所谓 (/users/845813)

上一篇: 会议论文汇总2:Recsys2020_39篇长文汇总 (/art...

1F 银汉 (/users/873599)

2020-10-30 18:58:58




点赞

所谓 赞同

 1 (/comments/311077/voteup)

 0

写下你的评论...



评论