

# 简历

## 个人资料

姓名：杨鹏  
出生日期：1987  
性别：男  
学历：本科，2006 - 2010，石家庄经济学院  
专业：计算机科学与技术  
电话：15652301070  
E-mail: yanggpeng@hotmail.com

## 工作经历

2011/03 - 2012/03：  
北龙中网 | 软件工程师  
高性能DNS服务器的设计和研发。

2012/03 - 2013/03：  
58同城 | 算法工程师  
数据挖掘，个性化推荐算法研发。

2013/03 - 2017/03：  
世纪佳缘 | 算法组组长  
带领7人算法团队，为各业务提供数据和算法支持，包括推荐系统研发，nlp，图像处理。

2017/03 - ：  
腾讯新闻 | 算法工程师  
负责腾讯新闻客户端主timeline的Ranking工作。

## 项目

腾讯新闻，技术栈切换到Tensorflow：  
算法从fm升级到deepfm，使用padding-then-mask和分布式训练等方法解决Tensorflow训练效率问题。自研线上inference方案代替tf-serving，解决线上预测超时的问题。线上效果提升3%。

腾讯新闻，改进mmoe优化mtl模型效果：  
模型从ctr单目标改为多目标，优化用户停留时长。基础结构使用mmoe，使用异构网络解决mmoe中expert学习只依赖随机初始化的问题，并改进gate函数，解决gate与expert依赖度低的问题。时长提升2%。

腾讯新闻，listwise重排曝光item：  
在现有推荐构架下增加refine层，利用self-attention建模item之间的影响，重新排序最终的item。线上效果提升5%。

腾讯新闻，知识蒸馏用于粗排：

推荐框架增加粗排模块，解决精排越来越复杂，线上耗时扛不住的问题。与精排模型联合训练，使用知识蒸馏提升效果。auc比精排低0.01，效率是精排的12倍。

腾讯新闻，改进fm算法：

使用点击历史学习用户表达，替换fm的uid特征，解决uid过于稀疏和fm不好建模相似用户的问题。线上效果提升2%，同时可以加速模型收敛，减少小流量实验观察时间。

世纪佳缘，推荐相关：

基于二部图的好友推荐算法，使用Spark GraphX实现，运行时间从4-5小时(mr实现)减少到20分钟。使用一种简单有效的方法解决单个用户发信或者收信太多导致的图节点爆炸问题。

发信读信点击预测模型，利用hadoop，hive分析佳缘用户行为数据，提取用户行为特征，使用lr，svd，gbdt等机器学习算法对用户点击进行预测，提高用户活跃度。测试模型融合，持续优化效果。

实时预测算法平台，基于dubbo开发，实现以上提到的模型，并提供分布式的实时服务。可以极大的提高模型测试及上线的效率。

世纪佳缘，nlp相关：

客服自动问答，对用户的问题进行分类后，聊天使用闲聊模式，业务问题基于ES实现。后期发现业务问题效果不好，重新基于分类和索引结合实现，回答正确率有很大提高。

聊天系统，使用word2vec提取聊天数据的语义，使用聚类算法划分场景。对新聊天对话使用分类算法划分场景并产生回复。

端到端的对话系统研发，正在进行的研究性项目，主要使用深度学习技术：cnn提取对话特征，rnn做对话状态追踪，cnn做意图识别，lstm做自然语言生成。

世纪佳缘，图像处理相关：

验证码识别，利用深度学习识别变长验证码。参考Google识别街景门牌号的论文，使用multi-task learning同时学习验证码长度和验证码数值。基于tensorflow实现的cnn模型，线上正确率98.5%。

颜值系统，在Google Inception-V3模型上进行fine-tuning训练，识别美女照片，把模型用于用户颜值评分。修改slim源码，离线训练模型，并提供线上实时服务。

生活照联系方式识别，识别用户上传生活照中嵌入的QQ号，微信号，手机号，网址等联系方式。首先尝试了Faster R-CNN，SSD识别所有文字，但是召回太低。最后采用CTPN论文中专门针对文字识别改进的模型，解决文字识别的召回问题，并以识别出来的文字框作为数据，训练一个cnn二分类模型识别是不是联系方式。

58团购推荐系统研发：

实现icf,ucf等召回算法。使用mpi, Hadoop做并行计算，利用WAND算法缩短计算时间。优化lr排序模型。

DNS递归服务器开发：

开发类似于opendns，支持客户自定义规则的dns递归服务器。使用bind做后台服务器，使用redis存储用户数据和规则。