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: yphoho@163.com

工作经历

2011/03 – 2012/03：
北龙中网 | 软件工程师
实现以机器学习算法为核心的入侵检测系统；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 – ：
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，金融风控等。

主要业绩

组建了一个工作能力很强的算法团队，积极跟进普及学术界和工业界最新的研究成果，应用在业务中，为提高算法效果不断努力；
研发上线的推荐算法排序模型提升用户活跃度，发信量相较之前有40%提升，点击提升90%以上；开发的客服自动回复系统，反作弊系统，极大的提高了客服部门的工作效率，减少了用户投诉；还有为VIP部门提供的排序算法，为无线提供的聊天系统都有很好的反馈；
积极的在组内学习推广深度学习，并使用在nlp和图像相关的项目上，取得了很好的效果；

项目

基于机器学习的入侵检测：
分析流量数据，使用Weka中的分类算法识别网络中僵尸主机，优化模型特征，提高检测率。

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。

58团购推荐系统研发：
实现youtube论文中提到的item-based cf算法，并以常用的user-based cf算法作为补充。在实现user-based cf时，在机器资源有限且计算量很大的情况下，使用mpi, Hadoop做并行计算，并使用WAND算法大大缩短计算时间。后处理阶段使用ctr预测排序进一步提高效果。

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。
yige.ai，使用机器学习，nlp等技术开发的chatbot聊天框架，方便开发者创建基于自然语言的用户接口。带领团队负责算法相关部分的开发工作。网址：www.yige.ai。
基于cnn的情感分析，使用word2vec提前训练的embedding初始化，在训练模型时使用现在的embedding和对embedding继续训练两个channel。优化网络结构，调整参数，效果最后比lr提高很多。
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。
鉴黄，识别用户上传的色情，不雅动作，异性合照等不合格照片。使用Inception模型，转化为多分类问题。使用各种image augmentation方法增加训练数据，调参优化模型效果。
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。