张文韬

150-1095-2913 | zhangwt@ihep.ac.cn

教育经历

中国科学院大学 2016年9月 - 2019年7月

计算机技术 硕士 计算机与控制学院 北京

- GPA：3.6 / 4.0
- 荣誉/奖项：学业奖学金（2016-2018）
- 相关课程：统计机器学习理论，高级人工智能，大数据系统与大规模数据分析，深度学习

中北大学 2011年9月 - 2015年7月

电子信息工程 本科 信息与通信工程学院 太原

- GPA：3.4 / 4.0

项目经历

基于强化学习的分布式存储系统的性能调优 2017年12月 - 2018年7月

- 分布式存储系统存在大量可供调节的参数，这些参数的设置对系统的性能有着很大的影响。而由于庞大的参数规模，负载的多样性以及调节和反馈之间的延时，人工调节往往存在偏差且效率低下。
- 如果把调节引擎看作是智能体，把存储系统看作是环境，存储系统的参数调节问题是典型的顺序决策问题，因此，项目基于强化学习实现了自动化的参数调优。
- 本人参与了项目的设计工作，并负责项目的全部编码实现工作。
- 项目使用的强化学习算法有DQN, A2C, PPO。以 lustre 分布式文件系统作为测试环境，以系统吞吐率作为目标性能，以 lustre 文件系统默认参数为基准，该方法可使系统吞吐率提升30%左右。

集群监控系统 2017年9月 - 2017年11月

- 存储系统往往存在冷热文件分布不适的问题导致了系统的利用率较差，计算中心因此需要建立文件级的存储监控系统，提高系统利用率的同时，对存储系统的使用情况也可以有更全面的了解。
- 监控系统分数据采集，数据处理以及数据可视化三部分。本人主要负责数据采集工作，并且参与了数据可视化的工作。
- python内置的扫描函数应用在分布式存储系统上速度很慢，因此实现了多线程并发扫描的算法，使采集速率提升数倍。基于各个维度的可视化工作使中心对存储系统的使用情况有了更深入的了解，并对存储系统的管理维护提供支持。

智能运维-时间序列的预测以及异常检测 2017年11月 - 2018年5月

- 时间序列的分析应用十分广泛。在计算中心的生产系统中，将时间序列预测应用于存储容量规划，基于时间序列预测，做一些统计学的分析将时间序列异常检测应用于集群节点的异常检测。
- 本人参与了存储容量的规划，并负责集群节点的异常检测。
- 调研相关论文，使用prophet来进行时间序列预测，该方法属于广义加性模型，在考虑趋势的同时，将季节性和假期作为两个因素考虑进来。异常检测方面，为了解决异常趋势的检测，将一定时间窗口内的预测差值建模为概率分布，并通过计算短时间窗口内的差值均值在该分布函数中的值来判断是否异常。

竞赛经历

kaggle-Mercari价格预测挑战

- 日本最大的社区购物网站Mercari希望为卖家提供定价建议,因为卖家可以在Mercari上投放任何东西，种类多，数量大，因此人工产品定价就非常困难。所以需要建立一个模型，自动建议正确的产品价格。数据为产品的文本说明，包括产品名称，类别，商标和详细的产品描述等信息。
- 使用TfidfVectorizer对数据进行预处理，将文本数据转换为数值型数据。基准模型为含3个隐藏层的全连接神经网络，交叉验证进行训练。基准模型的表现不够好，排名较低。
- 通过分析训练误差和验证误差，做了相应的针对欠拟合、过拟合的调整措施，并在最后使用了集成的方式，完成这些后排名相应提升到top1%。

论文发表

- HPC China 2018: Performance Optimization of Lustre File System Based on Reinforcement Learning (EI)
- CHEP 2018: Smart Analysis on EOS File storage

技能/证书及其他

- **技能：** Python，Java
- **语言：** 英语（CET-6）
- **兴趣爱好：** 篮球