**张智鹏**

**性别: 男 出生日期: 1993年9月**

**联系方式: (+86)15201199242 zhangzhipeng@pku.edu.cn**

**地址: 北京大学, 中国北京市海淀区颐和园路5号, 100871**

**教育背景**

**博士(北京大学) 计算机软件与理论 2015年9月-2020年6月**

* 研究兴趣: 大数据处理与分析, 分布式机器学习.
* 导师：崔斌 教授

**学士(山东大学) 计算机科学与技术 2011年9月-2015年6月**

* 排名1/15, 泰山学堂.

**学术/工作经历**

**高级算法工程师(阿里) 2020年7月-目前**

* 基于Flink的分布式机器学习算法，资源预估，分布式图嵌入算法.

**算法工程师(实习生, 腾讯) 2018年11月-2019年12月**

* 分布式机器学习算法，分布式图嵌入算法.

**访问交流(系统组, 苏黎世联邦理工学院) 2017年7月-2018年1月**

* 加速Spark上的机器学习.
* 导师：张策 教授

**开源项目**

**Apache Flink**

* Flink ML的联合发起人和维护者。Apache Flink committer.

**研究项目**

* **MLlib\*:用Spark快速训练广义线性模型 【发表于ICDE 2019，第一作者】**

Spark是业界最流行的大数据处理框架之一。据统计，在XX部门超过80%的数据使用Spark进行数据处理，却很少使用Spark的官方学习库MLlib，而是使用很多定制化的机器学习系统比如TensorFlow，Angel，Petuum等。这导致了分布式环境下巨大的数据移动和机器学习模型开发维护的代价。在这项工作中：

* + 详细分析了Spark MLlib和参数服务器系统的系统架构和实现，并且通过甘特图分析了Spark MLlib在执行SGD过程中的两个性能瓶颈：1）模型更新方式低效，2）通信方式低效。
  + 利用模型平均（model averaging）和AllReduce通信两个技术有效的解决了以上两个性能瓶颈。
  + 以无侵入的方式在Spark MLlib上实现了这两个技术（MLlib\*），并且从实验角度验证了MLlib\*的高效性：1）MLlib\*比MLlib快2-3个量级；2）MLlib\*能达到与参数服务器系统类似的性能。
  + 将模型平均和AllReduce两个技术拓展到了主题模型上，并且提供了理论证明和实验对比。实验表明MLlib\*能比Angel-LDA\*快5倍。
* **ColumnSGD:基于列划分的分布式SGD实现框架 【发表于ICDE 2020，第一作者】**

传统分布式机器学习系统基于行划分数据，在高维场景中面临存储和通信的巨大挑战。ColumnSGD将数据和模型按列划分到计算节点上，通过发送统计信息来计算梯度和更新模型。ColumnSGD在逻辑回归，因子分解机等上比Spark快930倍，比Petuum快63倍，比MXnet快14倍。

* **Spark之上的参数服务器系统PS2 【发表于SIGMOD 2019，第一作者】**

已有机器学习系统难以高效处理大数据。在此项目中，我们在Spark之上架设了参数服务器系统，并且提供了一个新的数据抽象DCV来管理机器学习模型。PS2比Spark MLlib快55.6倍，比Petuum快3.3倍。【注：此工作为与腾讯同事合作完成】

* **加速分布式图嵌入算法**

已有的基于参数服务器的分布式图嵌入算法通常会带来很大的通信开销或者括展性问题。在此项目中，我们将模型部分放到计算节点上，并且利用图划分的技术来减少节点之间的通信开销；进一步的，我们考虑用staleness减少网络带宽的使用。

* **SimRank算法的实验性比较 【发表于VLDB 2017，第一作者】**

SimRank是一种衡量图上节点相似性的算法，然而目前的解法多种多样，难以选择。我们将已有的算法进行一个分类，并且理论上给出了各种算法之间的联系，用复杂的实验说明了各种算法在不同情况下的优劣性。

**已发表论文**

1. **\*Zhipeng Zhang**, \*Yunyan Guo, Wentao Wu, Jiawei Jiang, Ce zhang, Bin Cui, Jianzhong Li. Model Averaging in Distributed Machine Learning: A Case Study with Apache Spark [**VLDBJ** 2021, **CCF-A**]
2. **Zhipeng Zhang**, Wentao Wu, Jiawei Jiang, Lele Yu, Bin Cui, Ce Zhang. ColumnSGD: A Column-oriented Framework for Distributed Stochastic Gradient Descent [**ICDE** 2020, **CCF-A**]
3. **Zhipeng Zhang**, Bin Cui, Yingxia Shao, Lele Yu, Jiawei Jiang, Xupeng Miao. PS2: Parameter Server on Spark [**SIGMOD** 2019, **CCF-A**]
4. **Zhipeng Zhang**, Jiawei Jiang, Wentao Wu, Ce Zhang, Lele Yu, Bin Cui. MLlib\*: Fast Training of GLMs using Spark MLlib [**ICDE** 2019, **CCF-A**]
5. **Zhipeng Zhang**, Yingxia Shao, Bin Cui, Ce Zhang. An Experimental Evaluation of SimRank-based Similarity Search Algorithms [**VLDB** 2017, **CCF-A**]
6. Wentao Zhang, Bin Yuan, **Zhipeng Zhang**, Bin Cui. Distributed Optimization and Implementation of Graph Embedding Algorithms. [**JOS 2021, CCF-A**]
7. Jiawei Jiang, Pin Xiao, Lele Yu, Xiaosen Li, Jiefeng Cheng, Xupeng Miao, **Zhipeng Zhang**, Bin Cui. PSGraph: How Tencent trains extremely large-scale graphs with Spark? [**ICDE** 2020, **CCF-A**]
8. Xiaoru Qu, Zhao Li, Jialin Wang, **Zhipeng Zhang**, …, Jun Gao. Category-aware Graph Neural Networks for Improving E-commerce Review Helpfulness Prediction. [**CIKM 2020, CCF-B**]
9. Xinyi Zhang, **Zhipeng Zhang**, Bin Cui. An Reinforcement Learning-based Method for Join Optimization. [**NDBC** 2020，**CCF-C, Best Student Paper**]
10. **Zhipeng Zhang**, Jiawei Jiang, Lele Yu, Bin Cui. Angel+: A Large-Scale Machine Learning Platform on Angel. [**Frontiers of Data and Computing, 2019**]
11. Jiawei Jiang, **Zhipeng Zhang**, Bin Cui, Yunhai Tong, Ning Xu. StroMAX: Partitioning-based Scheduler for Real-time Stream Processing System [**DASFAA** 2017, **CCF-B**]
12. Shuyang Shi, **Zhipeng Zhang**, Bin Cui. Resume Activeness Prediction in Online Recruitment Scenarios [**JCST** 2017, **CCF-B**]

**获得荣誉**

校长奖学金(top 2%) 校级 2016

搜狐奖学金(top 10%) 院级 2016

校长奖学金(top 2%) 校级 2017

学术品德奖(top 10%) 校级 2017

学术报告奖(top 2%) 院级 2018

秒针奖学金(top 10%) 院级 2019

北京大学二等奖学金（top5%） 校级 2019

**技能掌握**

**专业技能**

★★★★★Flink, Spark, Angel, Parameter Servers; Java, Scala; latex, shell, git.

★★★★ TensorFlow, Petuum, MXNet; C/C++, Python; yarn, docker.

**语言技能**

★★★★★ English

**自我评价**

专注细节，团队意识，责任心强，自我驱动.