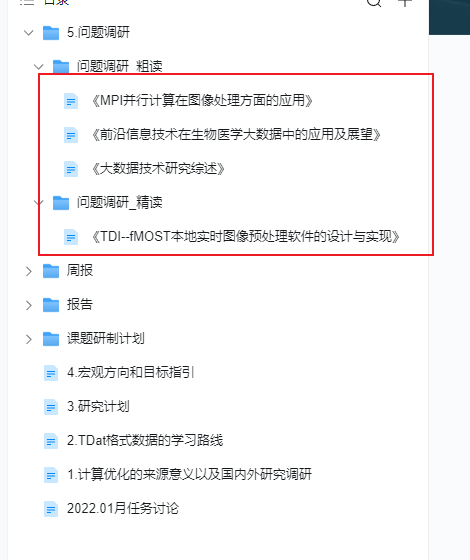
**周 报**

杨世城 21级专硕 2022年2月27日

|  |  |
| --- | --- |
| **课题任务简表** | |
| **课题任务** | 基于TDat数据的计算优化 |
| **近期目标** | 1. **大数据方向**——文献调研查阅资料，并输出问题记录文档； 2. 确立进一步具体工作方向； |
| **研究内容** | 1. 阅读《TDI--fMOST本地实时图像预处理软件的设计与实现》，输出笔记文档； 2. 阅读《大数据技术研究综述》，输出笔记文档； 3. 阅读《前沿信息技术在生物医学大数据中的应用及展望》，输出笔记文档； 4. 阅读《MPI并行计算在图像处理方面的应用》，输出笔记文档； 5. 查找资料，浏览文献； |
| **技术路线** | 1. 文献阅读，输出阅读文档； 2. 提炼问题，整理问题和解决方案笔记； |

**本周工作概要**

1. 整理阅读笔记和问题概览，输出文档知识库；



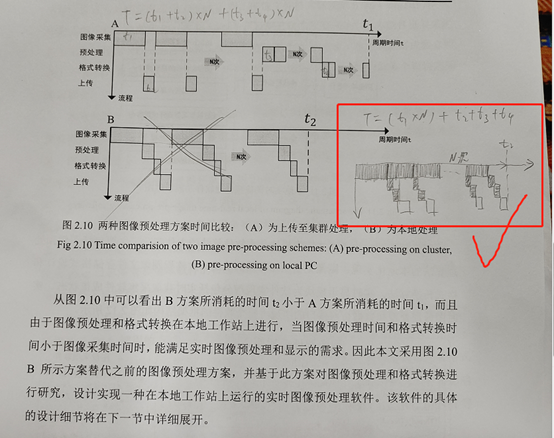
**1、《TDI--fMOST本地实时图像预处理软件的设计与实现》**

**核心启示：**

**当前**的处理流程：1.数据采集，2.数据上传，（全部数据采集完了之后），3.预处理，4.格式转换；

**理想**中可以转换流程改进为1.数据采集，（采集了一个单元之后，即可开始后续），2.预处理，3.格式转换，4.上传；

把分布在不同机器上的多任务进行转移到本地来执行，并按照流水线的结构进行组合，使之能够大大减少时间耗费；（详细时间计算可见图中）



**=========================**

**问题需求描述：**

图像数据的查看和分析滞后与数据的采集过程；

“然而对于脑神经科学研究人员，在采集过程中实时查看成像结果十分重要，例如获取病毒注射位点、神经元的投射和神经元突触结构的信息等，并且TDI-fMOST系统成像参数也需要根据成像效果实时调整。”

**问题处理思路：**

研究现状中对当前已有问题进行了描述，分为了三部分进行问题描述：“预处理”、“大体积格式转化”、“自适应分配节点的并行加速”。

**1.预处理**：

“由于无法自动化获取成像参数，这些软件无法在采集时实时进行图像预处理，需要等采集完毕后再进行预处理，这也增加了获取用于图像显示和分析数据的总时间。”

**解决思路：——调整处理顺序，以及处理单元。**

**2.大体积格式转换**：

“TDat作为一种通用三维图像数据处理平台，其需要图像数据预处理完毕后才能在集群中进行格式化处理。”

**“**TDat格式转换的基本单元为512×512×512像素的三维图像块，这就需要等待512层图像数据采集完毕，然而采集512层数据需要数个小时时间，这种方式无法满足实时处理的需求。**”**

**解决思路：——该文结合TDI-fMOST系统按冠状面逐层图像采集的特点，设计了一种以二维冠状面图像为基本单元的大体积图像数据存储格式。图像预处理后的二维冠状面图像经过图像分块和图像金字塔处理，产生多分辨率的图像块，并利用TIFF文件格式进行存储和索引。**

**3.并行加速**：

“然而在实际采集过程中，数据采集速度并不保持一致，TDI-fMOST系统通常会随着鼠脑样本区域动态改变数据采集速度，为了保证实时处理的需求，数据处理速度也需要随之改变。上述两种并行加速的方式没有考虑动态修改并行数目的情况，为满足实时处理的需求，通常会设置一个较大的并行数目，这种方式会造成硬件资源的不合理分配，甚至会造成数据采集丢帧、数据采集速度变慢等情况，影响图像采集软件。”

**解决思路：——线程池。（工程问题）**

**本文展望**

1. 是否可以在图象预处理算法上进一步增加对数据的精简，比如只对样本区域进行保留；
2. 图象预处理与图像采集当前还是独立开来的两个软件，所以之间会存在有一次数据的读取，后续是否可以把二者过程进行结合，从而减少这一次的数据从硬盘读取；

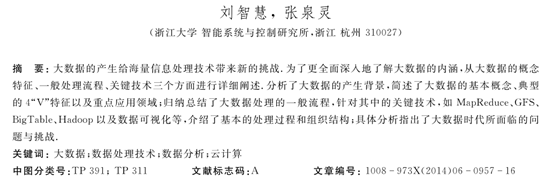
**2、《大数据技术研究综述》**

## 总启示：

优化分布式文件系统角度进行改进；

对文件存储方式从文件夹格式转为分布式并行数据库角度进行改进；

**====================================================**

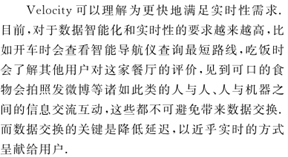


## 大数据的特征（4“V”）：

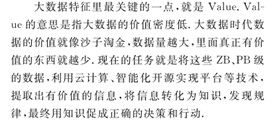
1.Volume（容量大），

2.Variety（种类多），

3.Velocity（速度快），可以理解为更快的满足实时性需求；



4.Vlaue（价值密度低），知识密度随着体量的变大而相对会变低，强调提取存储数据中有价值的那块；



## 大数据处理流程（该文总结）：（参考当前我们已有流程进行对照）

              数据采集——数据处理与集成——数据分析——数据解释

## 大数据关键技术：（参考当前我们已有流程进行对照）

              云计算（暂无，是否有应用？）

              分布式文件系统（我们当前用的Lustre）

              分布式并行数据库（对于当前以文件存储的形式是否有改进启发？）

~~大数据可视化（暂不涉及）~~

## 大数据带来的挑战：

~~安全与隐私（不涉及）~~

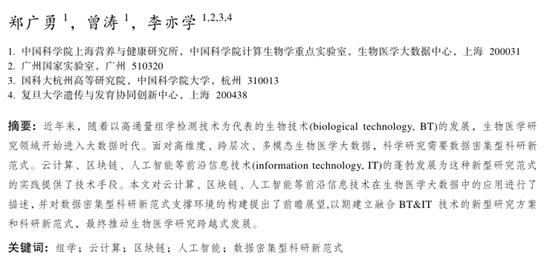
        大数据的集成与管理问题（压缩，清洗，存储）

        大数据的IT技术架构问题（不涉及）

~~大数据的生态环境问题（暂不涉及）~~

**3、《前沿信息技术在生物医学大数据中的应用及展望》**

## 《Application and prospect of cutting-edge information technology in biomedical big data》（即《前沿信息技术在生物医学大数据中的应用及展望》）



1. 云计算技术在生物医学大数据中的应用；（暂不涉及）“云计算(cloudcomputing)是分布式计算的一种，指的是通过网络“云”将巨大的数据计算处理程序分解成无数个小程序，然后通过多部服务器组成的系统进行处理和分析这些小程序得到结果并返回给用户。”.
2. 区块链技术在生物医学大数据中的应用；（暂不涉及）“在生物医疗领域迫切需要建立安全、互利的数据共享机制，从而使数据通过流通与汇聚释放价值，推进生物医药产业的创新发展。区块链技术的去中心化、可追溯、不可伪造、公开透明等属性赋予该技术应用于生物医疗领域数据管理共享的能力。”

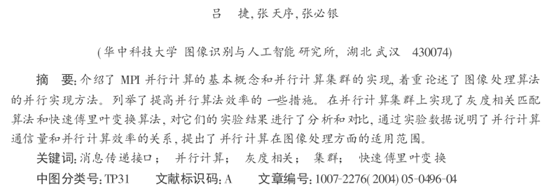
3. 人工智能技术在生物医学大数据中的应用；（已在图像识别，数据分析中应用）

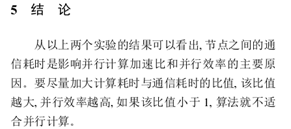
**4、《MPI并行计算在图像处理方面的应用》**

## 总启示：

可以从并行计算的角度进行优化改进和调研搜索

======================================================





1. 继续搜索大数据相关论文，发现问题继续阅读；

