2020-02-14

- 时间：21:30～23:00
- 与会人员：孙逸伦 田晨江 赵文祺 陈耿阳
- 记录人：田晨江

---

0. 会前准备
    1. 上次会议遗留问题
        - 数据处理与爬取是否限制为Java？**建议**使用Java。（已确认）
        - 过程中产生的文档可能需要满足IEEE规范。具体是什么规范？（已询问fcr，等待回复）
    2. 分工
        - 需求：cgy
        - 数据处理：zwq
        - Jenkins：syl
        - 代码覆盖率+搜索：tcj
    3. 螺旋模型

> （以下来自《软件工程与计算二》和Ron Patton的《软件测试》）
>
> 原型能够澄清不确定性，所以原型能够解决风险。所以，螺旋模型**使用原型**解决项目中从需求到设计的各种**风险**。
>
> **步骤**
>
> 1. 确定目标、可选方案和限制条件
> 2. 明确并化解风险
> 3. 评估可选方案
> 4. 当前阶段开发和测试
> 5. 计划下一阶段
> 6. 确定进入下一阶段的方法
>
> **特征**
>
> 1. 基本思想是尽早解决比较高的风险，如果有些问题实在无法解决，那么早发现比项目结束时再发现要好，至少损失要小得多。
> 2. 迭代与瀑布的结合，开发阶段是瀑布式的，风险分析是迭代的。
> 3. “风险驱动”
>
> **优点**
>
> 降低风险，减少因风险造成的损失。
>
> **缺点**
>
> 1. 使用原型方法，本身就有风险。
> 2. 模型过于复杂，不利于管理者依据其组织软件开发活动。

> **应用**
>
> 高风险的项目。

可以预见，这次项目需要拿出原型，无论是抛弃式的，还是演化式的。

```
4. c-dblp: http://cdblp.ruc.edu.cn/computer/scholarexplorer/
```

---

1. 已完成的部分

- 对原有数据集进行扩展
  - 原本2015-2019，现ASE 2000-2020 ICSE 2010-2020
  - 新增metrics属性
- 其他
  - 改写爬虫脚本为python3
  - 以Authors属性为依据进行数据清洗
  - 准备使用MongoDB、Json进行初步存储
  - IEEE速度较慢，5s一个，大规模爬取费时

2. 需要讨论的地方

- 其他可能的数据源：AAAI, IJCL, CVPR, NeurIPS
- references属性常常会连接到不在本数据库中的内容
- 如何判别不是论文的文章?
- 数据库类型的选择？关系型、非关系型

3. 下阶段任务

- 完成项目启动文档的草稿
- 对现有爬虫方法的优化，探索是否有更高效率的爬取方法
- 试验四种其他可能的数据源
- 由于项目开发采用螺旋模型，迭代一的时候先不宜过度考虑后续迭代的相关事项，以防止需求发生较大变动