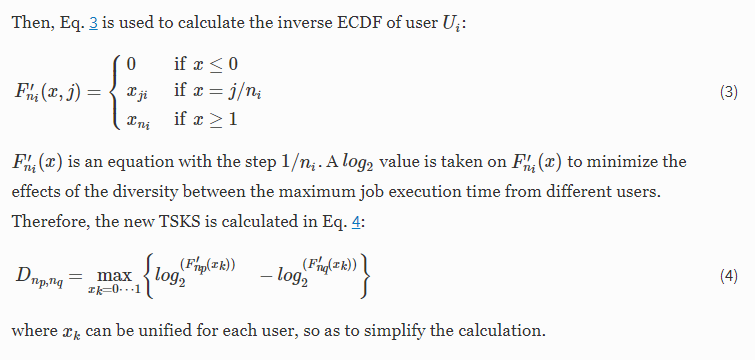
北航实验复现

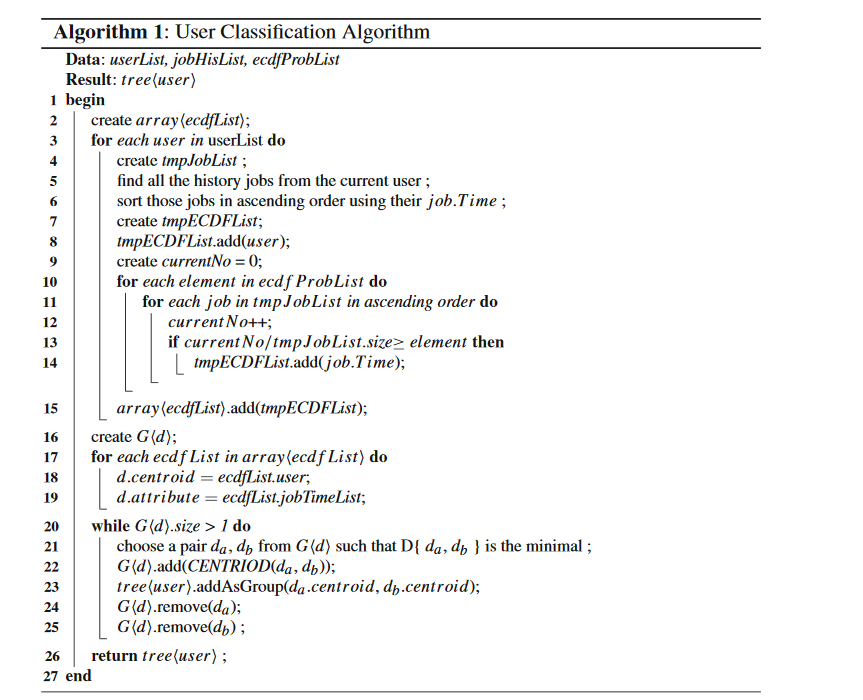
1. 背景
2. 目的：对网格计算中的用户进行聚类，根据聚类结果预测任务的执行时间
3. 数据来源：<http://gwa.ewi.tudelft.nl/datasets/gwa-t-1-das2>
4. 数据格式：保存在Sqlite数据库
5. 复现论文：A parallel job execution time estimation approach based on user submission patterns within computational grids
6. 原文链接：<https://link.springer.com/article/10.1007/s10766-013-0294-1>
7. 环境
8. 编程环境：Pycharm 2016.3.2
9. 硬件环境： Window 10
10. python版本：3.6.3
11. 项目位置：supercomputer->user\_cluster
12. 复现过程
13. 数据清洗
    * 1. 下载 Sqlite Developer
      2. 在Sqlite Developer 中附加在数据来源中下载的数据库
      3. 执行以下的SQL语句：
         1. Delete from jobs where status = 0 or runtime = 0;
         2. Select count(\*) from jobs;
         3. Select distinct userid from jobs;

注：2)和3)是为了将数据结果与论文中的结果进行比对。比对之后发现结果一致。任务数确实为1104921，用户数也确实为331。

1. 用户聚类算法实现
   * 1. 定义用户任务累积分布函数和定义用户相似度



* + 1. 用户聚类算法过程图：



* + 1. 代码
    2. 各参数定义：
       1. ecdfProlist = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]
       2. userList 可以从数据库中获取
       3. jobHistoryList 可以从数据库中获取
    3. 结果保存：F:/usermodel/datasource/beihang/user\_cluster.txt

1. 思考
   1. 它在进行用户聚类的过程中，只用到了任务的运行时间作为用户的提交模式的衡量标准。没有利用其它额外的特征。这样聚类出来的用户模型太过于单薄。
   2. 它在用户聚类的过程中，没有利用到任务时间上的关联性，没有把数据看做是时序数据。它仅仅是将每个用户的所有任务的运行时间按照运行时间长短进行排序之后，得到结果。
   3. 它在获取用户之间进行相似度比较的数据时，是根据预先定义好的ecdfprolist，但是这个向量的值的由来，并没有依据。且作者在论文中也没有提及。但是这个向量其实对于结果的影响是很大的。它直接决定了每个用户的数据粒度。
   4. 可以改进的地方：
      1. 在聚类的过程中加入更多的特征，比如用户所在的实验室
      2. 加入任务数据的时序特征
      3. 自动化确定ecdfprolist最好的取值