第六届“泰迪杯”
数据分析技能赛

优
秀
作
品

作品名称：档案数字化加工流程数据分析

荣获奖项：一等奖并获泰迪杯

作品单位：浙江工商大学

作品成员：卢茜　唐至阳　王苗

档案数字化加工流程数据分析

摘 要：

随着扫描、光学字符识别（OCR）、数码摄影、数据库、多媒体和存储技术的不断创新，档案数字化作为一种崭新的档案信息处理技术应运而生。它将各类载体的档案资源转化为数字形式，并以数字形态进行储存，实现网络化连接，通过计算机系统进行综合管理，筑构出一个有机有序的档案信息库。眼下，我国各个行业的存量档案数量巨大，对档案数字化的需求持续攀升，档案数字化加工行业的市场规模呈现逐年扩大之势。

本文基于某档案数字化加工单位 2020 年 7 月加工处理过程中各个工序的管理数据，对档案数字化流程的耗时和进度情况、操作人员的工作量和工作效率情况进行了统计分析与可视化展示，旨在为管理人员能够及时了解档案加工处理动态提供数据参考。具体做法如下：

针对问题一，数据预处理与统计。本文首先统计出完成四道工序的案卷数量为 33980，并在有效工作时间范围内将各案卷各工序的开始时间及各案卷的完成时长进行统计。其次，本文统计出需要返工的案卷数量为 8166，其占完工案卷总数的百分比为 24.032%，并将返工案卷的返工工序和返工开始时间进行汇总。同时，本文一方面对自检全检工序，汇总每个操作人员的返工案卷数，计算其占该操作人员该工序工作总量的百分比，另一方面按工序分别统计完成案卷的数量、总耗时和平均耗时。最后，本文按操作人员，根据不同工序，统计其工作时长、完成案卷的数量与每个案卷的平均耗时。

针对问题二，数据分析与可视化。本文根据给出的管理数据，绘制出多幅数据可视图，分别为每天不同工序完成案卷数量的簇状柱形图、各工序每天投入工作量（单位：人·小时）的多重折线图、每天各工序返工案卷数占当天返工案卷总数的百分比堆积面积图与每个操作人员返工案卷数饼图。

针对问题三，领取提交模式分析。本文通过可视化的方法分析批次内案卷的领取提交时序，总结有 4 种领取提交模式，分别为串行领取串行提交模式、多次领取多次提交、多次领取同时提交与同时领取多次提交模式，并对每一种模式提供了一个实际案例。

关键词：档案数字化，数据分析，数据可视化，案卷领取提交模式

目 录

一、问题简介

1.1 问题背景

随着扫描、光学字符识别（OCR）、数码摄影、数据库、多媒体和存储技术的不断发展，档案数字化作为一种新型档案信息处理技术应运而生。它将各种载体的档案资源转变为数字化档案信息，并以数字形式存储，以网络化形式相互连接，通过计算机系统进行管理，构建一个有序结构的档案信息库。我国档案工作采取“存量数字化、增量电子化”的信息化战略。当前，我国各行业存量档案数量巨大，档案数字化的需求持续增加，档案数字化加工行业的市场规模呈逐年增长之势。

1.2 问题重述

（一）预期目标

对加工流程数据进行统计分析，并作可视化展示，便于管理人员及时了解档案加工处理动态。具体目标如下：

1. 统计档案数字化流程的耗时和进度情况。
2. 统计操作人员的工作量和工作效率情况。

（二）附件内容

表 1-data.xlsx

表 2-result1_1.xlsx

表 3-result1_2.xlsx

表 4-result1_3.xlsx

表 5-result1_4.xlsx

表 6-result1_5.xlsx

表 7-result3.xlsx

注：data.xlsx 记录了某档案数字化加工单位 2020 年 7 月加工处理过程中各个工序的管理数据。

（三）任务要求

基于上述问题背景与提供的附件数据，本文需要研究完成以下任务：

任务一：数据预处理与统计

a)统计完成四道工序的案卷数量，在报告中列出统计结果。汇总各案卷各工序的开始时间及各案卷的完成时长，以表 1 的格式将汇总结果保存到文件“result1_1.xlsx”中，同时在报告中列出案卷完成时长最长的三个案卷的结果。

注 1：每个案卷的完成时长是扫描、图像处理、自检全检三个工序的耗时之和，PDF 处理无需计算耗时，各工序的耗时是该工序的开始时间至结束时间的时长。

注 2：完成时长应去掉非工作时间（“三、案卷加工流程说明”第 5 条），单位：h，保留 3 位小数。

b）统计需要返工的案卷数量及其占完工案卷总数的百分比，在报告中列出结果。汇总返工案卷的返工工序和返工开始时间，以表 2 的格式将汇总结果保存到文件“result1_2.xlsx”中，同时在报告中列出返工案卷号“托 40606-册六”“托 40606-册七”“托 5901_1-册三”的结果。

注：未返工工序的时间为空。

c）对自检全检工序，汇总每个操作人员的返工案卷数，计算其占该操作人员该工序工作总量的百分比，按百分比降序排列，以表 3 的格式将结果保存到文件“result1_3.xlsx”中，同时在报告中列出前三位操作人员的结果。结果保留 3 位小数，例如：返工案卷占比为 1%，在结果表中填写“1.000”。

d）按工序分别统计完成案卷的数量、总耗时和平均耗时，以表 4 的格式将结果保存到文件“result1_4.xlsx”中，并在报告中列出结果。结果保留 3 位小数。

注：按工序计算总耗时，是该工序各个批次的案卷集最早开始时间至案卷集最晚结束时间之和，而不是各个案卷完成时长的总和。

e）按操作人员、工序统计工作时长、完成案卷的数量和每个案卷的平均耗时（h/卷），以表 5 的格式将结果按操作人员 ID 升序排列保存到文件“result1_5.xlsx”中，同时在正文中列出操作人员 ID“10”“33”“48”的结果。结果保留 3 位小数。

注：按操作人员、工序统计工作时长是按批进行的（“三、案卷加工流程说明”第 3 条），应去除非工作时间（“三、案卷加工流程说明”第 5 条）。

任务二：数据分析与可视化

a)计算并绘制每天不同工序完成案卷数量的簇状柱形图：x 轴表示时间，y 轴表示完成案卷的数量，用不同颜色标记不同工序。

b)计算并绘制各工序每天投入工作量（单位：人·小时）的多重折线图：x 轴表示时间，y 轴表示每天投入的工作量，用不同颜色标记不同工序。

c)绘制每天各工序返工案卷数占当天返工案卷总数的百分比堆积面积图：x 轴表示时间，y 轴表示百分比，用不同颜色标记不同工序。

d)对图像处理工序，汇总每个操作人员返工案卷数，计算其占该工序返工案卷总数的百分比，并按百分比进行排序，绘制饼图，其中排名第 10 位及以后的合并成一个扇区。

任务三：领取提交模式分析

根据文件 data.xlsx 的批次数据，通过可视化的方法分析批次内案卷的领取提交时序，总结有哪几种领取提交模式。对每一种模式给出一个实际例子，以表 6 的格式保存到文件“result3.xlsx”中，同时在报告中参照图 1 和图 2 的方式分别绘制两种不同的示意图。

二、任务一：数据预处理与统计

2.1 问题分析

通过观察题目给出的附件数据与任务要求，本文分析得出任务一存在以下两大难点：

难点一：判断案卷已完成的标准。根据题目要求可知，虽然题目中注明每个案卷的完成时长是扫描、图像处理、自检全检三个工序的耗时之和，PDF 处理无需计算耗时，但是在对完成案件进行计数时，需要确定每个案件在四道工序中均有结束时间。

难点二：每个案卷完成时长的计算方式。由于实际工作中工人会出现提前上岗或推迟下岗的情况，通过观察附件数据可知，如果仅仅以每个案卷的第三道工序结束时间减去第一道工序的开始时间作为每个案卷的完成时长，则会将非工作时间也会计入到完成时间中。因此，本文对附件数据中可能存在的所有情况进行总结，后续问题求解时对于一些特殊情况需要进行特殊处理，具体见表 1。

表 1 工序情况归纳表

序号	具体内容
情况 1	工序开始时间与结束时间均在同一周内，工序开始时间与结束时间均属于工作时间
情况 2	工序开始时间与结束时间均在同一周内，工序开始时间与结束时间均属于非工作时间
情况 3	工序开始时间与结束时间均在同一周内，工序开始时间属于工作时间，工序结束时间属于非工作时间
情况 4	工序开始时间与结束时间均在同一周内，工序开始时间属于非工作时间，工序结束时间属于工作时间
情况 5	工序开始时间与结束时间出现跨周的情况，工序开始时间与结束时间均属于工作时间
情况 6	工序开始时间与结束时间出现跨周的情况，工序开始时间与结束时间均属于非工作时间
情况 7	工序开始时间与结束时间出现跨周的情况，工序开始时间属于工作时间，工序结束时间属于非工作时间
情况 8	工序开始时间与结束时间出现跨周的情况，工序开始时间属于非工作时间，工序结束时间属于工作时间

2.2 第一小问求解

根据任务一的要求，针对第一小问：

首先，需要确定每个案卷在四道工序中是否均有结束时间，如果没有，则判定该案卷并未完成四道工序，如果有，则判定该案卷已完成四道工序，并对该案卷进行计数。

其次，需要判断所有完成四道工序的案卷中他们的开始时间与结束时间是否出现跨周的情况，如果出现跨周的情况，在计算完成时长时需要提前去除周日一天的时间，再计算有效的工作时间。

针对案卷中各个工序的开始时间或结束时间处于非工作时间，本文进行如下处理，具体见表 2。

表 2 非工作时间处理方法表

工序开始/结束时间	所处时间段	处理方法
开始时间	0:00-8:30	设为当天 8:30
开始时间	12:00-13:00	设为当天 13:00
开始时间	18:00-24:00	设为后一天 8:30
结束时间	0:00-8:30	设为前一天 18:00
结束时间	12:00-13:00	设为当天 12:00
结束时间	18:00-24:00	设为当天 18:00

最后，在完成对案卷各个工序非工作时间的处理后，根据情况去除中午午休的一小时时间与晚上 18:00 至后一天早上 8:30 前的下班时间，即可得到各案卷的有效完成时长。

根据上述解题思路，本文最终求得完成四道工序的案卷数量为 33980，案卷完成时长最长的三个案卷的结果如表 3 所示。

表 3 案卷完成时长最长的三个案卷的结果表

案卷号	扫描开始时间	扫描结束时间	图像处理开始时间	图像处理结束时间	自检全检开始时间	自检全检结束时间	PDF 处理开始时间	PDF 处理结束时间	完成时长
托	2020/7/	2020/7/1	2020/7/1	2020/7/1	2020/7/1	2020/7/1	2020/7/2	2020/7/20	132.8
610390	2	8	8	8	8	8	0	15:41:39	63
-册一	8:49:34	14:50:33	15:02:36	15:13:41	15:14:26	15:24:08	15:41:18		
托	2020/7/	2020/7/3	2020/7/3	2020/7/3	2020/7/8	2020/7/1	2020/7/1	2020/7/18	85.94
610362	2	8:23:45	14:14:12	14:30:17	10:23:20	7	8	9:48:11	7
-册一	8:49:12					11:23:16	9:32:52		
托	2020/7/	2020/7/2	2020/7/2	2020/7/2	2020/7/2	2020/7/2	2020/7/2	2020/7/21	67.25
682034	13	0	0	0	0	1	1	13:13:16	4
-册一	11:27:0	15:04:42	16:14:28	16:15:39	16:23:32	12:39:36	13:07:09		
	9								

2.3 第二小问求解

根据任务一的要求，针对第二小问：

首先，根据字段名为“dPROC_TIME”的值，判断各个案卷在各个工序上是否需要返工。

其次，对所有具有返工时间的数据根据“案卷号”去除重复项，统计需要返工的案卷数量。

最终，计算其占完工案卷总数的百分比，并将汇总结果保存到文件“result1_2.xlsx”中。

根据上述解题思路，求得需要返工的案卷数量为 8166，其占完工案卷总数的百分比为 24.032%。返工案卷号“托 40606-册六”“托 40606-册七”“托 5901_1-册三”的结果如表 4 所示。

表 4 “托 40606-册六”“托 40606-册七”“托 5901_1-册三”结果表

案卷号	扫描	图像处理	自检全检	PDF 处理
托 40606-册六			2020/7/7 15:27:05	2020/7/7 15:43:57
托 40606-册七			2020/7/7 15:32:29	2020/7/7 15:43:57
托 5901_1-册三		2020/7/13 10:38:18		

2.4 第三小问求解

根据任务一的要求，针对第三小问：

首先，需要对自检全检工序，根据返工开始时间，汇总每个操作人员的返工案卷数。

其次，计算每个操作人员的返工案卷数占该操作人员该工序工作总量的百分比，并将百分比按降序排列。

最后，以结果保存到文件“result1_3.xlsx”中。

根据上述解题思路，前三位操作人员的结果具体如表 5 所示。

表 5 前三位操作人员结果表

操作人员 ID	返工案卷占比（%）
8	100.000
17	13.043
42	2.147

2.5 第四小问求解

根据任务一的要求，针对第四小问：

首先，根据各个工序是否存在结束时间，分别统计各个工序完成案卷的数量。

其次，根据各个工序中各个批次的案卷集最早开始时间至案卷集最晚结束时间之和，计算各个工序的总耗时。

最后，用总耗时去除以每个工序完成案卷的数量，计算各个工序的平均耗时。

根据上述解题思路，各个工序计算结果如表 7 所示。

表 7 非工作时间处理方法表

工序	完成案卷的数量	总耗时（h）	平均耗时（h/卷）
扫描	33985	9160.930	0.270
图像处理	33985	3921.912	0.115
自检全检	33985	2368.813	0.070
PDF 处理	33980	921.944	0.027

2.6 第五小问求解

根据任务一的要求，针对第五小问：

首先，按操作人员，根据不同工序，统计其工作时长、完成案卷的数量与每个案卷的平均耗时。

其次，将结果按操作人员 ID 升序排列保存到文件“result1_5.xlsx”中。

最后，列出操作人员 ID“10”“33”“48”的结果。

根据上述解题思路，操作人员 ID“10”“33”“48”的结果如表 8 所示。

表 8 操作人员 ID“10”“33”“48”结果表

操作人员 ID	工序	工作时长（h）	完成案卷的数量	每个案卷的平均耗时
10	图像处理	5.665	9	0.629
10	扫描	5.665	9	0.629
10	自检全检	84765.245	11579	7.321
33	扫描	20277.45	2304	8.801
48	图像处理	28709.935	3632	7.905

三、任务二：数据分析与可视化

3.1 第一小问求解

根据任务二的要求，针对第一小问，结合每天不同工序完成案卷数量，本文得到的簇状柱形图如图 1、图 2、图 3、图 4 所示，具体数值如表 9、表 10、表 11、表 12 所示。

表 9 2020 年 7 月 1 日至 7 月 8 日数值表

时间	工序	完成案卷的数量	时间	工序	完成案卷的数量
7 月 1 日	扫描	523	7 月 2 日	扫描	883
7 月 1 日	图像处理	231	7 月 2 日	图像处理	557
7 月 1 日	自检全检	1	7 月 2 日	自检全检	100
7 月 1 日	PDF 处理	0	7 月 2 日	PDF 处理	1
7 月 3 日	扫描	1253	7 月 6 日	扫描	1460
7 月 3 日	图像处理	1089	7 月 6 日	图像处理	1105
7 月 3 日	自检全检	962	7 月 6 日	自检全检	664
7 月 3 日	PDF 处理	694	7 月 6 日	PDF 处理	906
7 月 7 日	扫描	1601	7 月 8 日	扫描	1519
7 月 7 日	图像处理	1224	7 月 8 日	图像处理	1049
7 月 7 日	自检全检	1284	7 月 8 日	自检全检	1034
7 月 7 日	PDF 处理	713	7 月 8 日	PDF 处理	1251

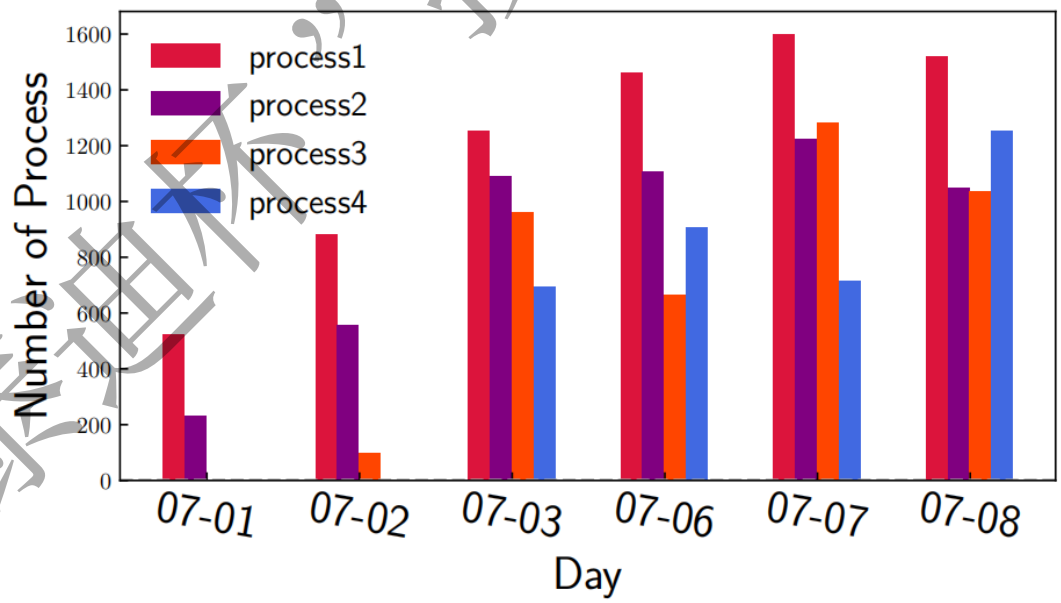

图 1 2020 年 7 月 1 日至 7 月 8 日簇状柱形图

（注：process1、process2、process3、process4 分别表示扫描、图像处理、自检全检、PDF 处理）

表 10 2020 年 7 月 9 日至 7 月 15 日数值表

时间	工序	完成案卷的数量	时间	工序	完成案卷的数量
7 月 9 日	扫描	1113	7 月 10 日	扫描	921
7 月 9 日	图像处理	1675	7 月 10 日	图像处理	1346
7 月 9 日	自检全检	1244	7 月 10 日	自检全检	1237
7 月 9 日	PDF 处理	632	7 月 10 日	PDF 处理	1914
7 月 11 日	扫描	1213	7 月 13 日	扫描	978
7 月 11 日	图像处理	1072	7 月 13 日	图像处理	1070
7 月 11 日	自检全检	898	7 月 13 日	自检全检	821
7 月 11 日	PDF 处理	698	7 月 13 日	PDF 处理	989
7 月 14 日	扫描	1469	7 月 15 日	扫描	1287
7 月 14 日	图像处理	1369	7 月 15 日	图像处理	1460
7 月 14 日	自检全检	1447	7 月 15 日	自检全检	1542
7 月 14 日	PDF 处理	1355	7 月 15 日	PDF 处理	1348

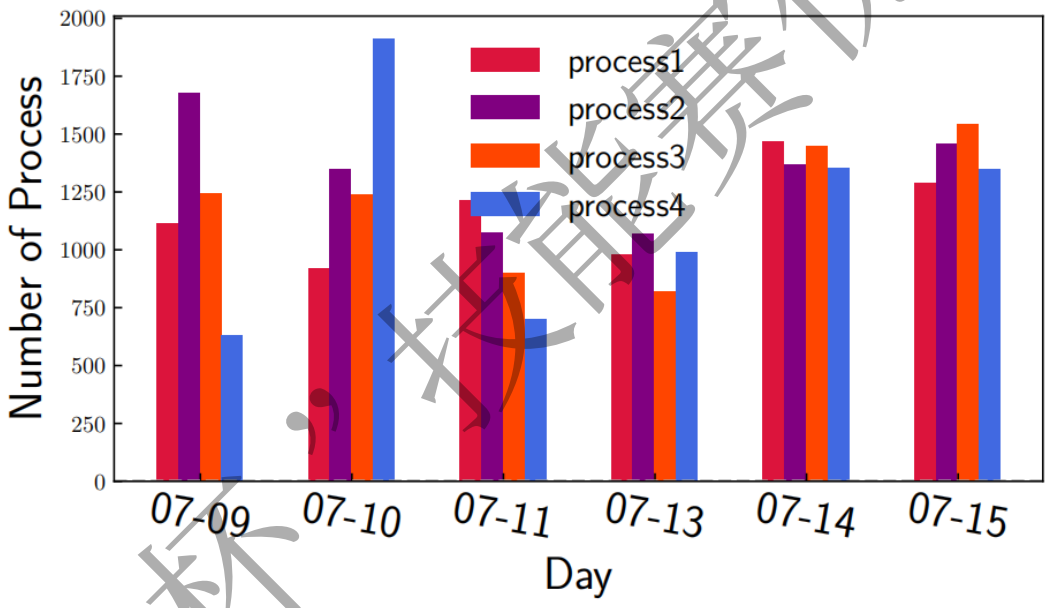

图 2 2020 年 7 月 9 日至 7 月 15 日簇状柱形图

（注：process1、process2、process3、process4 分别表示扫描、图像处理、自检全检、PDF 处理）

表 11 2020 年 7 月 16 日至 7 月 22 日数值表

时间	工序	完成案卷的数量	时间	工序	完成案卷的数量
7 月 16 日	扫描	2271	7 月 17 日	扫描	1722
7 月 16 日	图像处理	1390	7 月 17 日	图像处理	1228
7 月 16 日	自检全检	1368	7 月 17 日	自检全检	1155
7 月 16 日	PDF 处理	1801	7 月 17 日	PDF 处理	310
7 月 18 日	扫描	1823	7 月 20 日	扫描	1767
7 月 18 日	图像处理	1061	7 月 20 日	图像处理	1533
7 月 18 日	自检全检	1472	7 月 20 日	自检全检	1502

7 月 18 日	PDF 处理	1524	7 月 20 日	PDF 处理	2108
7 月 21 日	扫描	1661	7 月 22 日	扫描	2188
7 月 21 日	图像处理	1860	7 月 22 日	图像处理	1426
7 月 21 日	自检全检	2092	7 月 22 日	自检全检	1732
7 月 21 日	PDF 处理	1950	7 月 22 日	PDF 处理	1902

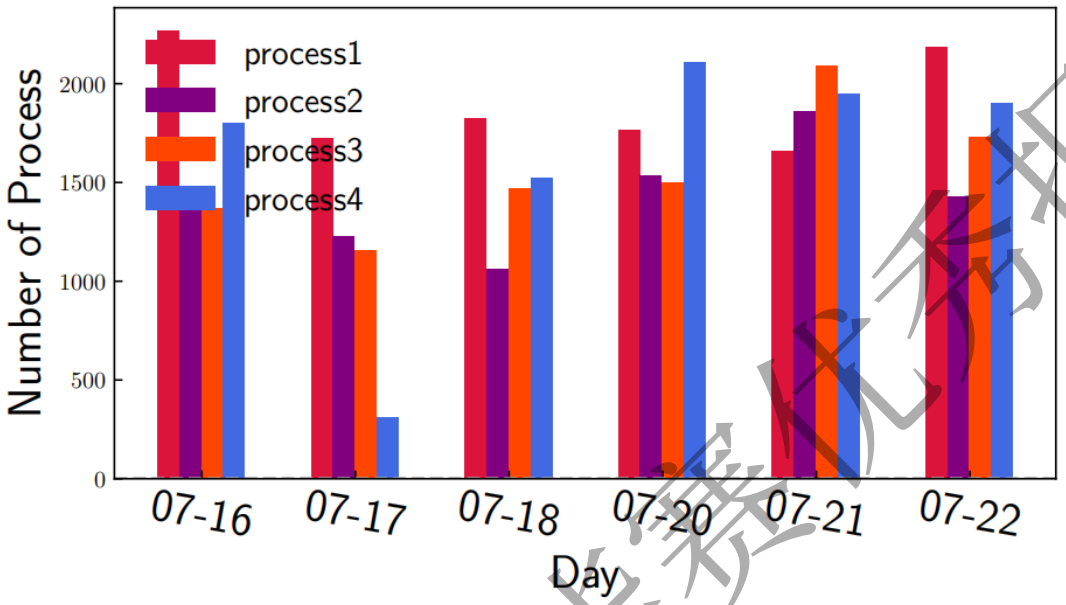

图 3 2020 年 7 月 16 日至 7 月 22 日簇状柱形图

（注：process1、process2、process3、process4 分别表示扫描、图像处理、自检全检、PDF 处理）

表 12 2020 年 7 月 23 日至 7 月 31 日数值表

时间	工序	完成案卷的数量	时间	工序	完成案卷的数量
7 月 23 日	扫描	1523	7 月 24 日	扫描	1352
7 月 23 日	图像处理	1543	7 月 24 日	图像处理	1880
7 月 23 日	自检全检	1705	7 月 24 日	自检全检	2068
7 月 23 日	PDF 处理	1804	7 月 24 日	PDF 处理	108
7 月 25 日	扫描	1129	7 月 27 日	扫描	1493
7 月 25 日	图像处理	1540	7 月 27 日	图像处理	1718
7 月 25 日	自检全检	1653	7 月 27 日	自检全检	1626
7 月 25 日	PDF 处理	3745	7 月 27 日	PDF 处理	1081
7 月 28 日	扫描	1218	7 月 29 日	扫描	946
7 月 28 日	图像处理	1496	7 月 29 日	图像处理	1626
7 月 28 日	自检全检	1975	7 月 29 日	自检全检	1470
7 月 28 日	PDF 处理	1407	7 月 29 日	PDF 处理	2502
7 月 30 日	扫描	672	7 月 31 日	扫描	0
7 月 30 日	图像处理	1730	7 月 31 日	图像处理	707
7 月 30 日	自检全检	1799	7 月 31 日	自检全检	1134
7 月 30 日	PDF 处理	1311	7 月 31 日	PDF 处理	1926

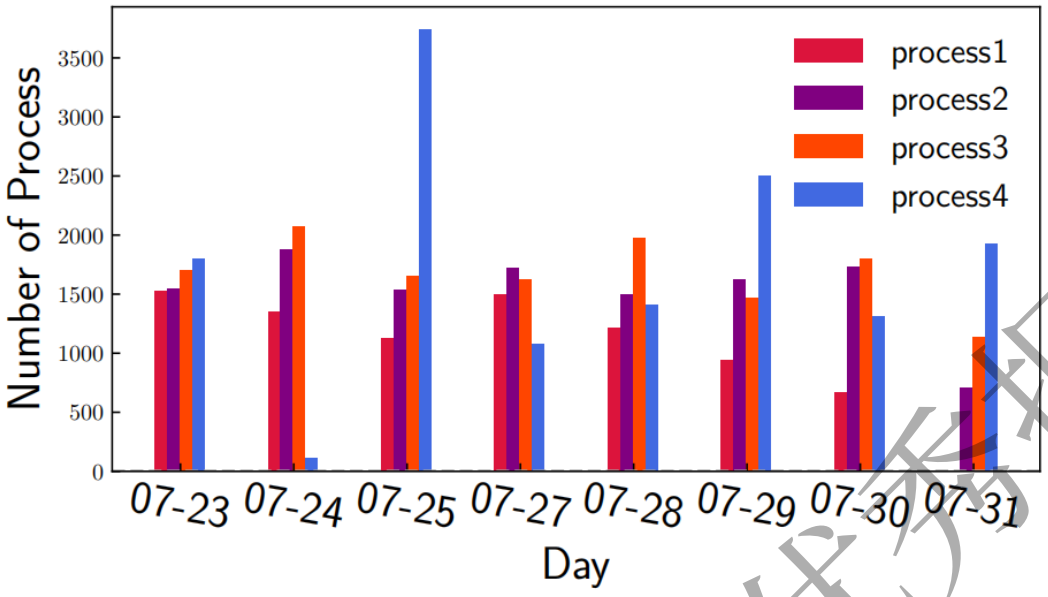

图 4 2020 年 7 月 23 日至 7 月 31 日簇状柱形图

（注：process1、process2、process3、process4 分别表示扫描、图像处理、自检全检、PDF 处理）

3.2 第二小问求解

根据任务二的要求，针对第二小问，结合各工序每天投入工作量，本文得到的多重折线图如图 5 所示。

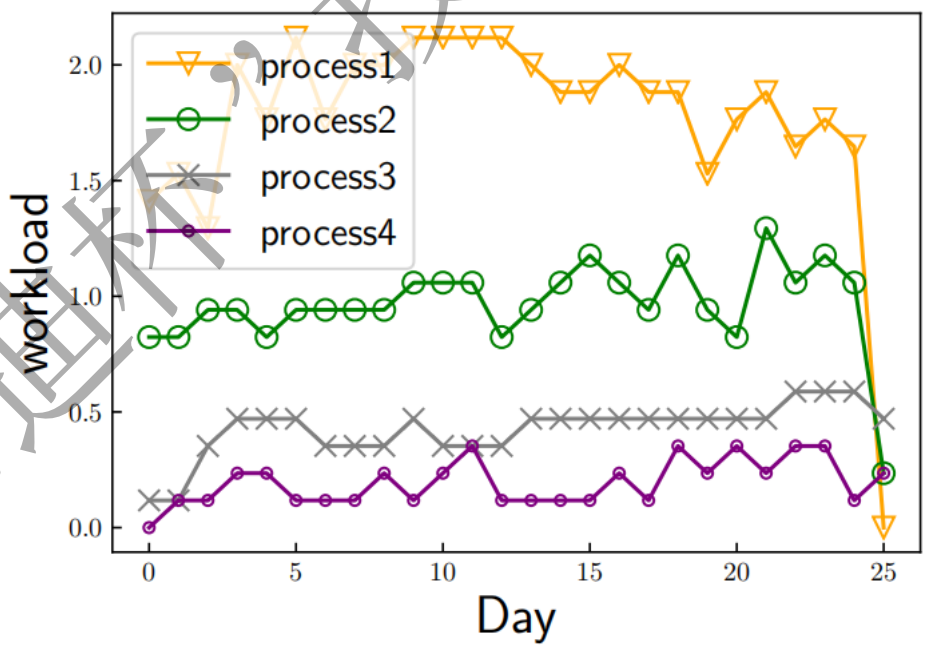

图 5 2020 年 7 月 23 日至 7 月 31 日多重折线图

（注：process1、process2、process3、process4 分别表示扫描、图像处理、自检全检、PDF 处理）

3.3 第三小问求解

根据任务二的要求，针对第三小问，结合每天各工序返工案卷数占当天返工案卷总数的百分比，本文得到的百分比堆积面积图如图 6 所示，具体数值如表 13 所示。

表 13 2020 年 7 月 1 日至 7 月 31 日每日各工序返工数值占比表

时间	工序	占比（%）	时间	工序	占比（%）
7 月 3 日	扫描	0.000	7 月 6 日	扫描	0.000
7 月 3 日	图像处理	100.000	7 月 6 日	图像处理	40.000
7 月 3 日	自检全检	0.000	7 月 6 日	自检全检	40.000
7 月 3 日	PDF 处理	0.000	7 月 6 日	PDF 处理	20.000
7 月 7 日	扫描	10.000	7 月 8 日	扫描	26.733
7 月 7 日	图像处理	0.000	7 月 8 日	图像处理	28.713
7 月 7 日	自检全检	50.000	7 月 8 日	自检全检	19.802
7 月 7 日	PDF 处理	40.000	7 月 8 日	PDF 处理	24.752
7 月 9 日	扫描	32.258	7 月 10 日	扫描	34.211
7 月 9 日	图像处理	37.097	7 月 10 日	图像处理	34.211
7 月 9 日	自检全检	17.742	7 月 10 日	自检全检	18.421
7 月 9 日	PDF 处理	12.903	7 月 10 日	PDF 处理	13.158
7 月 11 日	扫描	100.000	7 月 13 日	扫描	64.865
7 月 11 日	图像处理	0.000	7 月 13 日	图像处理	5.405
7 月 11 日	自检全检	0.000	7 月 13 日	自检全检	18.919
7 月 11 日	PDF 处理	0.000	7 月 13 日	PDF 处理	10.811
7 月 14 日	扫描	45.455	7 月 15 日	扫描	0.000
7 月 14 日	图像处理	0.000	7 月 15 日	图像处理	0.000
7 月 14 日	自检全检	27.273	7 月 15 日	自检全检	55.556
7 月 14 日	PDF 处理	27.273	7 月 15 日	PDF 处理	44.444
7 月 16 日	扫描	34.965	7 月 17 日	扫描	0.820
7 月 16 日	图像处理	1.399	7 月 17 日	图像处理	81.967
7 月 16 日	自检全检	54.545	7 月 17 日	自检全检	14.754
7 月 16 日	PDF 处理	9.091	7 月 17 日	PDF 处理	2.459
7 月 18 日	扫描	1.896	7 月 20 日	扫描	0.184
7 月 18 日	图像处理	97.156	7 月 20 日	图像处理	91.544
7 月 18 日	自检全检	0.948	7 月 20 日	自检全检	4.963
7 月 18 日	PDF 处理	0.000	7 月 20 日	PDF 处理	3.309
7 月 21 日	扫描	0.362	7 月 22 日	扫描	0.000
7 月 21 日	图像处理	74.275	7 月 22 日	图像处理	92.901
7 月 21 日	自检全检	14.130	7 月 22 日	自检全检	4.321
7 月 21 日	PDF 处理	11.232	7 月 22 日	PDF 处理	2.778
7 月 23 日	扫描	0.588	7 月 24 日	扫描	0.000
7 月 23 日	图像处理	98.627	7 月 24 日	图像处理	83.511
7 月 23 日	自检全检	0.588	7 月 24 日	自检全检	8.283

7 月 23 日	PDF 处理	0.196	7 月 24 日	PDF 处理	8.207
7 月 25 日	扫描	0.361	7 月 27 日	扫描	0.000
7 月 25 日	图像处理	97.115	7 月 27 日	图像处理	97.249
7 月 25 日	自检全检	2.524	7 月 27 日	自检全检	2.033
7 月 28 日	PDF 处理	0.000	7 月 27 日	PDF 处理	0.718
7 月 28 日	扫描	0.000	7 月 29 日	扫描	0.095
7 月 28 日	图像处理	96.759	7 月 29 日	图像处理	95.632
7 月 28 日	自检全检	2.593	7 月 29 日	自检全检	2.469
7 月 28 日	PDF 处理	0.648	7 月 29 日	PDF 处理	1.804
7 月 30 日	扫描	0.097	7 月 31 日	扫描	0.000
7 月 30 日	图像处理	97.182	7 月 31 日	图像处理	94.579
7 月 30 日	自检全检	1.944	7 月 31 日	自检全检	3.551
7 月 30 日	PDF 处理	0.777	7 月 31 日	PDF 处理	1.869

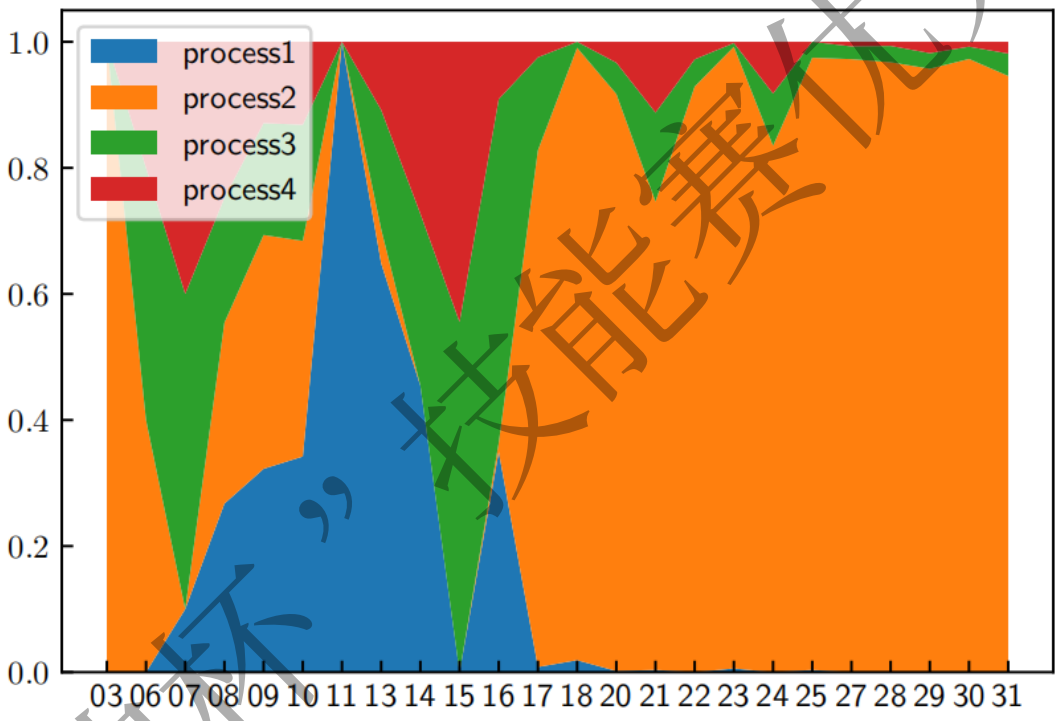

图 6 2020 年 7 月 1 日至 7 月 31 日每日各工序返工数值占比百分比堆积面积图

3.4 第四小问求解

根据任务二的要求，针对第四小问，本文观察 data.xlsx 发现，在图像处理工序中 7739 条返工数据有 7569 条数据尚未分配返工操作人员，仅有 170 条已分配返工操作人员，在本小问计算每个操作人员返工案卷数占该工序返工案卷总数的百分比中，本文将存在返工开始时间，但为分配返工操作人员的案卷不计入返工案卷总数中，在此思路下本文得到的饼图如图 7 所示，具体数值如表 14 所示。

表 14 前 10 位操作人员 ID 号及返工案卷所占百分比表

操作人员 ID	百分比（%）	操作人员 ID	百分比（%）
42	34.706	73	24.118
10	14.118	95	12.353
91	3.529	19	2.353
33	2.353	13	1.765
89	1.176	排名第 10 位及以后	3.529

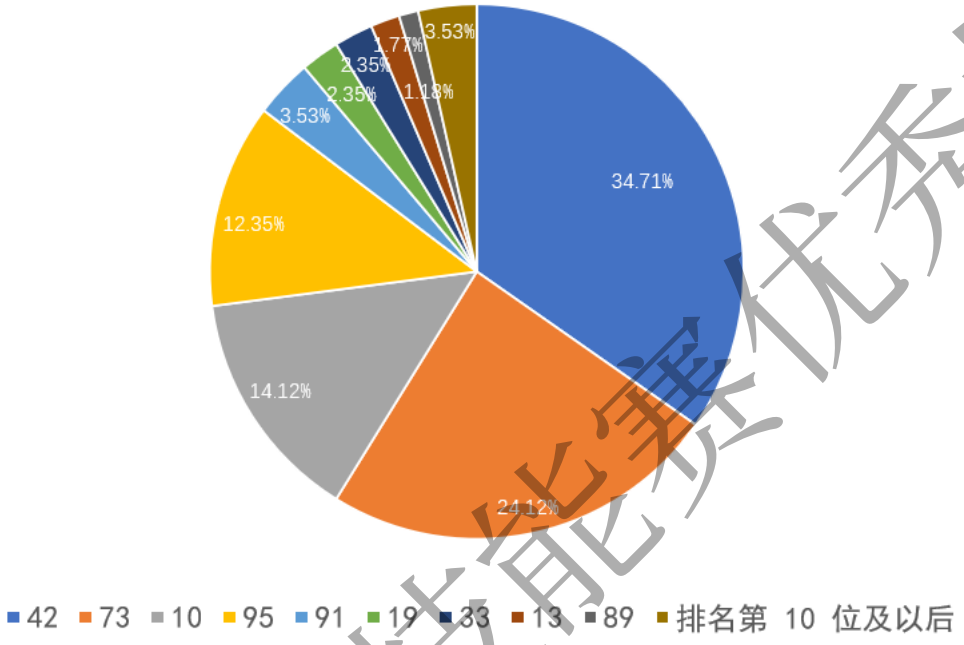

图 7 前 10 位操作人员 ID 号及返工案卷所占百分比饼图

四、任务三：领取提交模式分析

4.1 问题分析

本文根据操作人员 ID 以及批次编号对数据中案卷领取提交时序进行数据可视化，可以通过观察相同批号案卷操作时，操作人员领取和提交时间，以及同一操作人员领取提交连续案卷的时间等情况看出案卷领取提交模式，具体模式如下：

1、 串行领取串行提交：按顺序逐级领取和提交；

2、 多次领取多次提交：案卷可能经过多次领取和提交；

3、多次领取同时提交：案卷在不同时间领取但是在同一时间提交；

4、同时领取多次提交：案卷在同一时间领取但在不同时间提交。

4.2 问题求解

下面，本文对上述提到的 4 种模式都给出了一个实际例子，具体如图 8、图 9、图 10、图 11、图 12、图 13、图 14、图 15 所示。

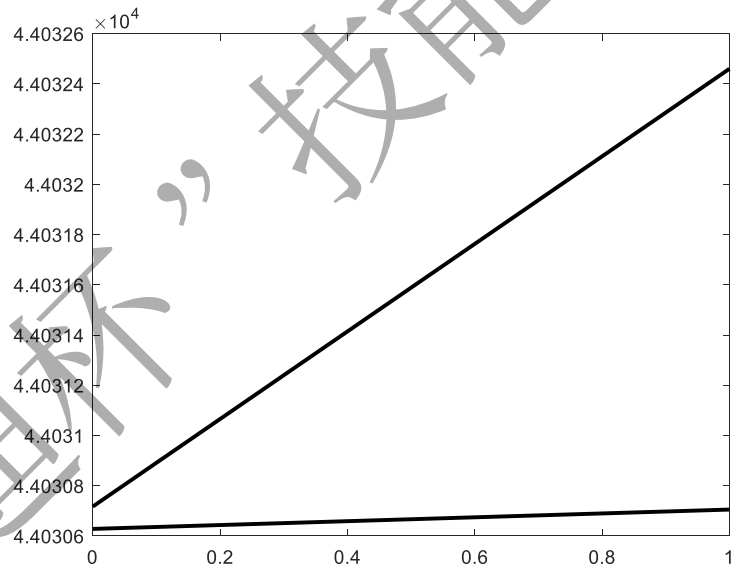

图 8 串行领取串行提交案例图

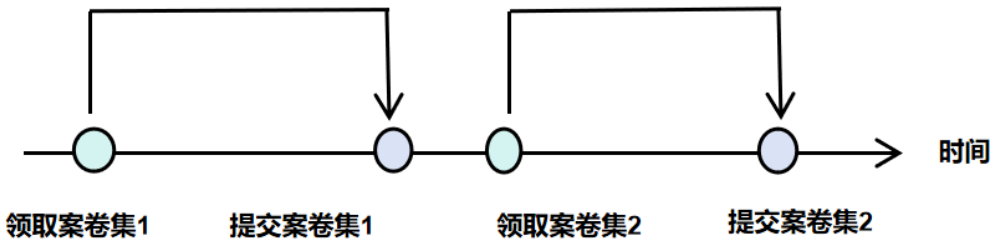

图 9 串行领取串行提交概念解释图

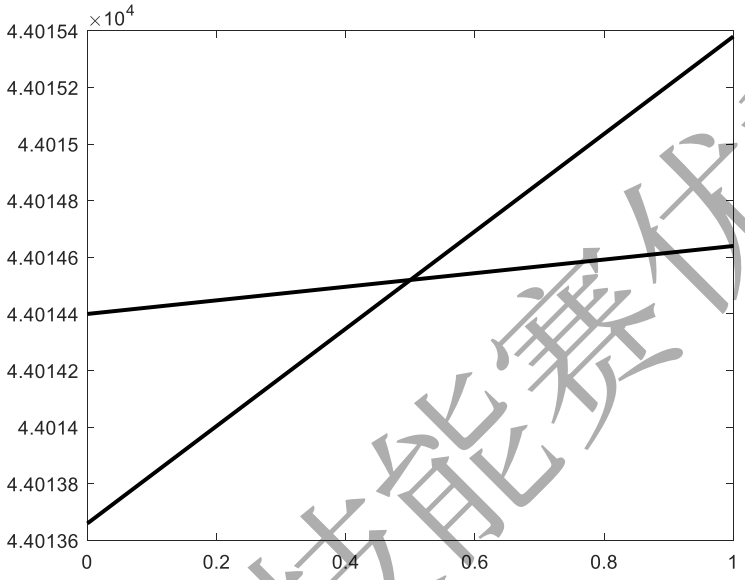

图 10 多次领取多次提交案例图

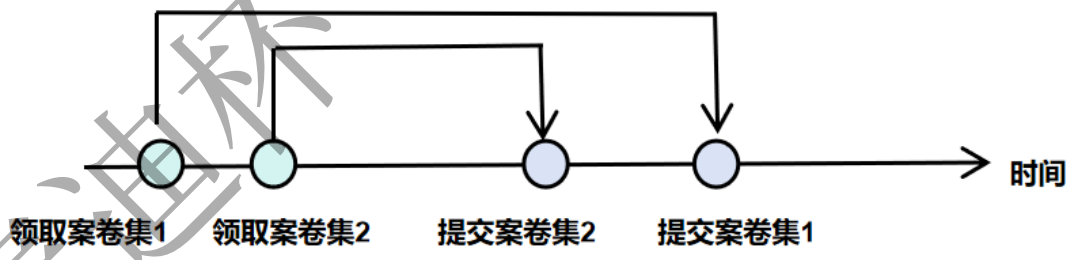

图 11 多次领取多次提交概念解释图

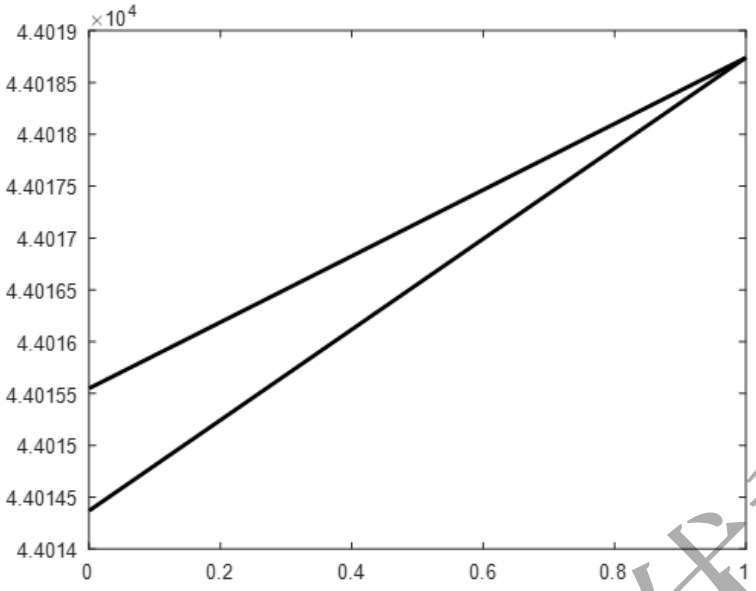

图 12 多次领取同时提交概念解释图

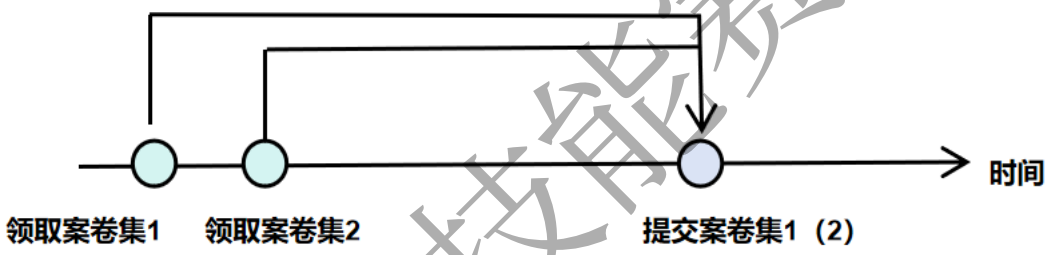

图 13 多次领取同时提交概念解释图

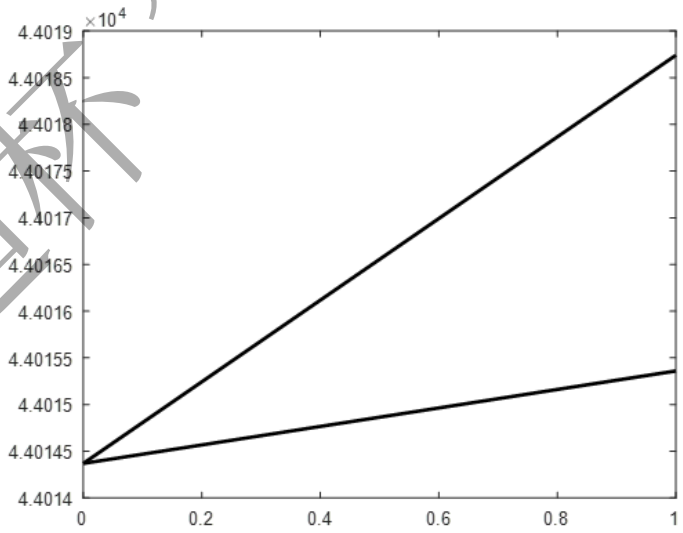

图 14 同时领取多次提交案例图

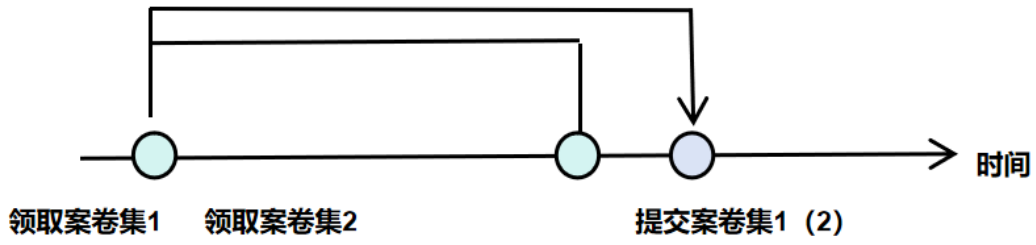

图 15 同时领取多次提交概念解释图

附录-相关代码