

**基于Scrapy的模板化爬虫程序管理平台**

**工作量统计分析报告**

v2.0

赵正阳 SY1906428

郭浩隆 SY1906430

沈一聪 SY1906510

梁远志 SY1906503

宋冰晨 SY1906429

2020年6月5日

**版本变更记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本 | 日期 | 修改人 | 修改章节 | 备注 |
| 1.0 | 2020/3/30 | 赵正阳 | 全文 | 完成初稿、统计实验准备和实验1的数据 |
| 1.1 | 2020/4/24 | 赵正阳 | 1.3、1.7、1.9 | 完成实验2、实验6和实验8截至第8周的数据统计 |
| 1.2 | 2020/5/8 | 赵正阳 | 1.4、1.7、1.9 | 完成实验3、实验6和实验8截至第10周的数据统计 |
| 1.3 | 2020/5/18 | 赵正阳 | 1.4、1.5、1.7、1.9 | 完成实验3、4、6、8截至第12周的数据统计 |
| 1.4 | 2020/5/29 | 赵正阳 | 1.5、1.6、1.7、1.9 | 完成实验4、5、6、8截至第13周的数据统计 |
| 2.0 | 2020/6/4 | 赵正阳 | 全文 | 增加引言，完善内容，完成总结 |

**目录**

[1. 引言 1](#_Toc42260705)

[1.1 编写目的 1](#_Toc42260706)

[1.2 工作量统计说明 1](#_Toc42260707)

[2. 工作量统计 1](#_Toc42260708)

[2.1 实验准备 1](#_Toc42260709)

[2.2 实验1软件需求分析 3](#_Toc42260710)

[2.3 实验2软件需求评审 4](#_Toc42260711)

[2.4 实验3软件设计与实现 5](#_Toc42260712)

[2.5 实验4软件测试 7](#_Toc42260713)

[2.6 实验5软件测试评审 9](#_Toc42260714)

[2.7 实验6软件项目计划与监控 11](#_Toc42260715)

[2.8 实验7软件配置管理 11](#_Toc42260716)

[2.9 实验8软件工程实验追踪与分析 12](#_Toc42260717)

[2.10 实验总结 12](#_Toc42260718)

[3. 总结与经验 13](#_Toc42260719)

[3.1 总结 13](#_Toc42260720)

[3.2 经验 13](#_Toc42260721)

1. **引言**

1.1 编写目的

本文档用于统计基于Scrapy的模板化爬虫程序管理平台(EasySpider)在每个实验阶段各组员的工作量。通过对实验项目的跟踪，分析每个实验过程的执行情况和结果，对比各组员的工作量差异及原因分析，最后给出整体的分析与总结。

1.2 工作量统计说明

本项目组制定的组员贡献率计算方式在实验准备文档《H-EasySpider-项目计划书》第4节中给出。本文档统计各组员在每个实验中的客观工作量数据，主要包括文档字数和图表数、代码行数以及其他工时3类，这些统计数据将用于最终的贡献率计算。其中，其他工时为撰写文档和写代码以外的其他工作（例如文档整合），与文档字数和代码行数所代表的是不同的工作。

本文档统计的工作量仅包括各组员不同的工作（例如撰写文档的章节分工、开发任务的模块分工等），而所有组员相同的工作（例如参加小组会、制作ppt及课堂汇报等）已在工作日志及实验6文档《进度控制分析报告》中体现，本文档不计入统计。

本文档的所有统计数据将同步至Excel表格“H-EasySpider-贡献率统计.xlsx”，在该Excel表格中将工作量统计数据代入贡献率计算公式中，从而可以自动计算出各组员的贡献率。

1. **工作量统计**

2.1 实验准备

实验准备阶段的主要工作是明确目标、工作范围和重点，确定实验组成员、实验项目以及计划安排，输出文档为项目计划书。

表2.1 实验准备工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | 项目简介、组员介绍 | Web UI | 反爬虫 | 分布式 | 整体计划、贡献率计算公式 |
| 项目计划书完成字数/图表数 | 1236/1 | 184/3 | 2207/4 | 788/1 | 679/1 |
| 其他工作 | 初稿整合 |  |  |  |  |
| 工时 | 1 h |  |  |  |  |

实验准备阶段各组员的工作量对比如图2.1所示。

(a) 考虑其他工时

(b) 不考虑其他工时

图2.1 实验准备工作量对比

图2.1(a)显示赵正阳占了60%的工作量，这是因为赵正阳有整合文档的工作作为其他工时，而其他组员没有这一项。根据贡献率计算公式，这一项将转换为50%的贡献率，因此看起来工作量占了非常大的比例。图2.1(b)为将其他工时的因素去除后的工作量对比，即仅考虑文档的完成字数和图表数。可以看出，此时沈一聪的贡献率最大，为42%。这是由于沈一聪在文档中贡献了较多的字数和图表数，而其他组员的工作量较为均衡。

2.2 实验1软件需求分析

需求分析阶段的主要工作是实践软件需求的分析、建模与验证方法，确定软件需求。对我们的系统EasySpider以及所依赖的框架Scrapy和Scrapyd进行需求分析，绘制用例图和RUCM图。实验输入为选择的开源软件及相关资料，输出文档为软件需求规格说明书。

表2.2 需求分析工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | Scrapy框架的需求分析 | Web UI部分的RUCM图 | 后端及调度程序 | Scrapyd及分布式的需求分析 | Web UI部分的需求文字描述 |
| 需求个数 | 14 | 10 | 17 | 10 |  |
| 需求规格说明书完成字数/图表数 | 3020/16 | 1090/10 | 3032/9 | 2414/14 | 1854/0 |
| 其他工作 | 初稿整合 |  |  |  |  |
| 工时 | 2 h |  |  |  |  |

需求分析阶段各组员的工作量对比如图2.2所示。

(a) 考虑其他工时

(b) 不考虑其他工时

图2.2 需求分析工作量对比

与实验准备类似，由于赵正阳有整合文档工作作为其他工时，因此图2.2(a)显示赵正阳占的大部分的工作量。而图2.2(b)去除其他工时后宋冰晨的贡献率仍然较小，这是因为宋冰晨在需求规格说明书中没有负责用例图及RUCM图的绘制，而其他组员的工作量较为均衡。

2.3 实验2软件需求评审

需求评审阶段的主要工作是实践软件评审，确定评审方法，确认软件需求。进行小组间的需求分析评审，填写需求评审检查单，并根据收到的意见进行需求修订和复评审。实验输入为软件需求规格说明书和相关资料，输出文档为需求评审报告。

由于该实验的主要工作量在于对被评审组提出评审意见，以及根据评审组提出的意见修改需求规格说明书，因此本实验的工作量由提出和修改的评审意见总数来衡量，每个评审意见折合0.2个工时。

表2.3 需求评审工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | 初评审G组  复评审A组  评审意见的修改 | 初评审G组  复评审I组  评审意见的修改 | 初评审F组  复评审A组  评审意见的修改 | 初评审F组  复评审I组  评审意见的修改 | 初评审G组  复评审I组  评审意见的修改 |
| 提出意见数 | 14 | 11 | 19 | 36 | 28 |
| 作出修改数 | 33 | 4 | 23 | 14 | 15 |
| 其他工作 |  |  |  |  |  |
| 工时 |  |  |  |  |  |

需求评审阶段各组员的工作量对比如图2.3所示。

图2.3 需求评审工作量对比

该实验的贡献率有较多的影响因素。一方面，提出的评审意见具有一定的主观性。组员提出的评审意见既有文字格式方面的也有用例设计方面的，文字格式问题较容易发现且数量多，用例设计方面的问题比较难以发现且数量少。提出这两类评审意见的工作量实际是不同的，但难以衡量具体的比例关系，在工作量统计时将所有的评审意见无差别对待，提出较多文字格式问题的组员将会比提出较少用例设计问题的组员“占优势”，这是导致最终反映出的贡献率不准确的原因之一。另一方面，实验1的分工会影响修改评审问题的工作量分配。由于收到的评审意见是根据需求规格说明书的分工分配到各组员，如果评审组对某组员负责的章节提出的评审意见少，则该组员修改的评审问题数也少。

以上两个问题导致该实验统计出的工作量不同组员差距较大，提出+修改的意见数最多的有50而最少的只有15。为了减少这种差距对最终贡献率的影响，在将意见个数折合为工时时会乘以一个较小的系数（如0.2）。

2.4 实验3软件设计与实现

软件设计与实现阶段的主要工作是在实验1~2的基础上设计并实现软件需求，进行编码和测试工作。实验输入为开源软件的当前版本和需求规格说明书，输出为实现方案文档和改进后的软件。

本项目选择的改进方式为“基于现有软件，开发新的应用软件”，即基于Scrapy框架开发一个模板化爬虫程序管理平台，而不是对Scrapy框架本身进行改进。因此最终需要提交的软件为一个Web项目，其中包含了使用Scrapy编写的爬虫程序模板供用户使用。

该实验的工作量统计中“代码行数”来自实验7文档《H-EasySpider-变更与管理分析报告》4.3节，是在所有开发工作完成后由梁远志使用Git命令统计得到。

表2.4 软件设计与实现工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | Django网站后端 | 登录注册和模板管理模块Django前端 | 爬虫模板 | 调度器 | 模板浏览、任务操作和数据下载模块Django前端 |
| 代码行数 | Python: 2208  HTML: 54 | Python: 9  HTML: 1268  CSS: 768  JavaScript: 405 | Python: 2280 | Python: 739  HTML: 311 | Python: 264  HTML: 2505  CSS: 350  JavaScript: 237 |
| 设计文档完成字数/图表数 | 1779/2 | 73/2 | 1897/1 | 1354/5 | 871/5 |
| 其他工作 | 整合初稿 |  |  |  |  |
| 工时 | 2 h |  |  |  |  |

软件设计与实现阶段各组员的工作量对比如图2.4所示。由于该实验最主要的工作量在于写代码，因此为考虑设计文档的整合工时。

图2.4 软件设计与实现工作量对比（不考虑其他工时）

根据本实验组在开发任务中的分工，宋冰晨和郭浩隆负责Django网站前端开发，因此主要的代码量为HTML, CSS和JavaScript；赵正阳负责Django网站后端，梁远志负责调度器组件，沈一聪负责爬虫模板的编写，因此主要的代码量为Python。而不同的语言开发难度不尽相同，例如Python代码比较精简，而HTML代码较多，因此在计算工作量时给每种语言的代码行数乘以一个难度系数。但每种语言的特点各不相同，很难给出一个绝对合理的难度系数，只能通过组员的主观判断，讨论商定。另外，本项目的整体架构设计、服务器部署等工作均由梁远志负责，该实验的工作量暂未统计这一部分的工作量，这是一个需要改进的地方。

2.5 实验4软件测试

软件测试阶段的主要工作是实践软件测试方法，开展软件测试。实验输入为被测软件，输出文档包括测试需求规格说明书、被测软件执行结果分析报告、软件问题报告等。

本项目组在该实验中使用了多种测试方法编写测试用例，包括单元测试、集成测试、场景测试、功能测试和非功能测试，也产生了较多的输出文档。表2.5中的“测试用例个数”仅为了统计各组员编写的测试用例数量，不计入工作量计算，因为每个测试用例已被计算为代码行数或测试需求规格说明书中对应的字数，因此不再重负计算。

表2.5 软件测试工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | Django单元测试、部分功能性测试、文档整合 | 部分功能性测试 | Scrapy单元测试、非功能性测试、性能测试、测试需求文档初稿及整合 | 调度器单元测试、部分功能性测试、集成测试 | 场景测试 |
| 测试用例个数 | Django单元测试：75  场景测试：1  功能测试：5 | 功能测试：7 | Scrapy单元测试：16  集成测试：1  非功能测试：11  性能测试：3 | 调度器单元测试：16  集成测试：2  功能测试：6 | 场景测试：5  功能测试：15 |
| 测试需求规格说明书完成字数 | 2327 | 2631 | 7178 | 4982 | 6742 |
| 软件测试计划完成字数/图表数 |  |  |  |  | 5519/3 |
| 被测软件执行结果分析报告完成字数 |  |  |  | 3705 | 78 |
| 软件可用性报告完成字数 |  |  |  |  | 3031 |
| 软件问题报告完成字数 |  |  |  | 1482 |  |
| 性能测试报告完成字数/图表数 |  |  | 883/2 |  |  |
| 其他工作 | 测试需求规格说明书v1.1整合 |  | 软件测试调研、测试需求规格说明书v1.0整合 | 列出测试用例和需求用例的对照表、整理撰写测试结果及问题报告 |  |
| 工时 | 3 h |  | 8 h | 5.5 h |  |

软件测试阶段各组员的工作量对比如图2.5所示。

图2.5 软件测试工作量对比

从图中可以看出，该实验中各组员的工作量比例非常不均衡，主要有两方面的原因。一方面，沈一聪和梁远志在实验初期进行了大量的调研和准备工作，因此在其他工时上占有较大比例。另一方面，该实验的产出文档较多，而宋冰晨负责了较多的文档编写工作，因此工作量也相对较大。而导致这些问题的根本原因是我作为组长在实验之前没有认真考虑工作量分配，导致部分组员的工作量过大或过小，属于我的工作失误。在实验5的工作分配时将更加注重工作量的均衡。

2.6 实验5软件测试评审

软件测试评审阶段的主要工作是采用规范的评审方法，确认测试需求、测试方法和测试结果。进行小组间的测试评审，填写测试评审检查单，并根据收到的意见进行修改和复评审。实验输入为被测软件以及测试需求规格说明书、被测软件执行结果分析报告、软件问题报告等实验4文档，输出文档为测试需求评审报告和软件问题报告。

与实验2相同，本实验也用提出和修改的评审意见总数来衡量工作量，每个评审意见折合0.2个工时。

表2.6 软件测试评审工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | 初评审E组  复评审C组  评审意见的修改 | 初评审E组  复评审C组  评审意见的修改 | 初评审D组  复评审B组  评审意见的修改 | 初评审D组  复评审B组  评审意见的修改 | 初评审D组  复评审C组  评审意见的修改 |
| 提出意见数 | 20 | 25 | 20 | 19 | 54 |
| 作出修改数 | 23 | 10 | 13 | 11 | 13 |
| 其他工作 |  | 整合评审意见 |  |  |  |
| 工时 |  | 3 h |  |  |  |

软件测试评审阶段各组员的工作量对比如图2.6所示。

图2.6 软件测试评审工作量对比

与实验2类似，实验3和实验4的分工也会影响到该实验的工作分配。但在分配修改评审问题工作时已考虑工作量的均衡性，因此与实验2相比，该实验各组员的工作量更加均衡。宋冰晨提出的评审意见较多，而郭浩隆有整合评审意见的额外工作，因此统计出的工作量比其他组员高一些，除此之外各组员之间的工作量没有太大差距。

2.7 实验6软件项目计划与监控

软件项目计划与监控部分的主要工作是软件开发的进度控制与任务协调。使用Microsoft Project进行进度控制并撰写进度控制分析报告。实验输入为影响因素分析和各组员每周的工作日志，输出文档为进度控制分析报告。该实验主要由宋冰晨负责。

表2.7 软件项目计划与监控工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | 提交工作日志 | 提交工作日志 | 提交工作日志 | 提交工作日志 | 维护MPP文件、进度控制分析报告、提交工作日志 |
| 进度控制分析报告完成字数/图表数 |  |  |  |  | 7086/10 |
| 工时 |  |  |  |  | 25.5 h |

2.8 实验7软件配置管理

软件配置管理部分的主要工作是在软件研发过程中，采用有效方法进行软件变更控制和版本管理。本项目组使用的配置管理工具为GitHub，过程要点包括配置管理计划、变更控制、版本管理、配置管理分析。实验输入为变更请求、更新版本，输出文档为变更与管理分析报告。该实验主要由梁远志负责。

表2.8 软件配置管理工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 |  |  |  | GitHub仓库管理、变更与管理分析报告 |  |
| 变更与管理分析报告完成字数 |  |  |  | 13755 |  |
| 其他工作 |  |  |  |  |  |
| 工时 |  |  |  |  |  |

2.9 实验8软件工程实验追踪与分析

软件工程实验追踪与分析部分的主要工作是通过对实验项目的跟踪，分析各项软件工程实验过程的执行情况和结果，比如工作量差异及其原因分析等。实验输入为各项软件实验过程统计数据，输出文档为工作量统计分析报告。该实验主要由赵正阳负责。

表2.9 软件工程实验追踪与分析工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 | 设计统计表格、每个实验的工作量统计、工作量统计分析报告 |  |  |  |  |
| 工作量统计分析报告完成字数/图表数 | 5918/8 |  |  |  |  |
| 其他工作 | 统计工作量数据 |  |  |  |  |
| 工时 | 5 h |  |  |  |  |

2.10 实验总结

实验总结的主要工作为实验1~8的实验内容和实验方法介绍、实验效果和问题分析、实验资料汇总以及综合实验效果与问题分析。实验输入为所有实验的制品及统计数据，输出文档为实验总结报告。

表2.10 实验总结工作量统计表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 姓名 | 赵正阳 | 郭浩隆 | 沈一聪 | 梁远志 | 宋冰晨 |
| 分工 |  |  |  |  |  |
| 实验总结报告完成字数 |  |  |  |  |  |
| 其他工作 |  |  |  |  |  |
| 工时 |  |  |  |  |  |

1. **总结与经验**

3.1 总结

综合以上所有统计数据，各组员总贡献率如图3.1所示。

图3.1 总贡献率对比

虽然各组员的总贡献率因为实验中的一些原因存在一定的不均衡，但也不存在过大的差距。从上文所总结的统计数据来看，每位组员尽其所能，从各自擅长的方向参与到本项目软件工程实验的各个环节中，最终保证了本项目的顺利完成。

3.2 经验

1. 明确的阶段性目标是推进项目进行的动力。在项目进行过程中，每周都会开1~2次小组会来安排下周的工作及分工。在开会之前确定一个主题，明确组会上要讨论的内容能够极大地提高讨论效率，每个组员有明确的任务分工也能够保证工作被高效地完成。
2. 组员的配合和老师的指导是项目顺利完成的重要保证。在项目初期，由于组员互相不熟悉，再加上本学期由于疫情只能采取远程会议的方式，组员之间沟通起来存在一定的障碍。经过一段时间的磨合，组员之间的默契程度逐渐提高。作为组长，仅靠我一个人的能力不可能完成整个项目的所有工作。组员对我工作的配合、出现问题时及时交流沟通以及老师对我们工作中存在的问题的指导都对项目的顺利完成起到至关重要的作用。
3. 开发工作只是软件工程的一小部分。起初我以为这门课只是要组队开发一个软件，但实际上写代码仅仅是其中的一小部分。其他的工作如需求分析、需求评审、软件测试等都是非常重要的环节。事实证明，我们在这些实验过程上花费的时间也远比开发工作要多。
4. 每位组员各有擅长的领域，组员之间互相学习、共同进步，在项目中积累团队协作经验是我们经过本项目最大的收获。