文章编号：2095-2163(2020)01-0307-05 中图分类号：311.5 文献标志码：A

GitHub 开源软件项目团队协作过程评价

刘玉辉，王忠杰
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：随着开源软件的兴起，为了得到更好的锻炼提升，计算机相关专业教师鼓励学生在 GitHub 上进行项目团队协作。针对学生在 GitHub 上协作完成的项目，教师如何进行项目团队内成员的贡献度量，进而为学生课程任务做出公平、公正的评分则成为一个问题。本文主要从构建成员贡献行为指标模型、设计量化贡献计算方法和成员贡献可视化等方面进行软件仓库挖掘，结合 SpringMVC、Hibernate 和 Extjs 设计并实现了一款 GitHub 团队项目成员贡献评估 Web 应用系统。通过对比实际项目人工评估结果和系统评估结果，验证了所提方法的有效性。

关键词：软件仓库挖掘；GitHub；团队协作；贡献；Web 应用

Mining GitHub for evaluating team collaboration process among OSS projects

LIU Yuhui, WANG Zhongjie
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the rise of open source software, computer-related professional teachers encourage students to collaborate on project teams on GitHub in order to get a better exercise upgrade. For projects on GitHub that students collaborate, how teachers can measure the contributions of members within the project team, and then the fair evaluation of student curriculum tasks become a problem.This paper focuses on software warehouse mining from the aspects of constructing member contribution behavior index model, designing quantitative contribution calculation method and visualization of member contribution. After that, the paper designs and implements a GitHub team project member contribution evaluation Web application system with Spring MVC, Hibernate and Extjs. Through specific experimental tests, the actual project manual evaluation results and system evaluation results are compared to verify the effectiveness of the proposed method.

【Key words】 software warehouse mining; GitHub; team collaboration; contribution; Web application

0 引 言

随着开源技术的兴起，GitHub 等开源软件取得了巨大的成功，目前，GitHub 上已经拥有超过 6 700 万个 repo（代码仓库），活跃用户达到 2 400 万人，超过 150 万家公司和机构进驻。在开源环境下，为了更好地实现团队合作，计算机相关专业教师也鼓励学生在 GitHub 上尝试项目开发协作，更好地锻炼提升自己。而针对学生在 GitHub 上协作完成的项目，教师如何进行项目团队内成员的贡献度量，进而为学生课程任务给出公平、公正的评分则成为一个问题。

传统的评分策略有：

（1）小组内 TeamLeader 根据小组成员平时的表现制定组内评级参考表，供教师或助教参考。

（2）教师或助教通过对组内成员提问考核，根据回答状况来核定评分等。

这些传统的方法在一定程度上可以解决项目团队成员评分问题，但同时也存在一些不足，如：没有充分利用项目团队开发过程的数据信息、主观因素会有部分影响、最终的问答形式考核趋于片面，作弊应付的可能性增大等。

为了解决上述问题，本文提出 GitHub 开源软件项目团队协作过程评价的研究，可以帮助教师解决 TA 问题。在 GitHub API v3 中共列举了 37 种事件，如当对一次提交进行评论时，就会触发 *CommitCommentEvent* 事件。这些事件中包括项目内部多人合作触发的事件和项目外部其它人员通过 Fork→PullRequest→Merge 途径做出贡献，本文主要研究项目内部多人合作时，各成员做出的贡献并进行度量。通过分析成员的贡献行为建立成员贡献行为指标模型，然后设计量化贡献计算方法，从而直观得出项目内成员的个人贡献。这样充分挖掘并利用了项目团队协作开发过程中在 GitHub 软件仓库中保留的数据信息，使项目团队内成员的贡献度量更

作者简介：刘玉辉（1992-），女，硕士研究生，主要研究方向：软件工程；王忠杰（1978-），男，博士，教授，博士生导师，主要研究方向：服务计算、软件工程。

收稿日期：2018-05-22

加准确。

1 相关工作

在软件项目中,如何对项目团队内开发人员的贡献度进行度量一直是软件仓库挖掘(Mining Software Repository,MSR 领域)广受关注的一个研究课题,传统的度量方法包括 Sackman 等人[1]提出的以单词量作为工作单位、程序的代码行数(Lines of Code,LOC)、面向对象中的类以及函数构件等,其中面向对象的软件开发中各种工作单位的讨论在文献[2]中已有清晰阐述。除了程序代码的贡献,还有人从项目团队成员之间的协作讨论、修复 bug 情况等方面相继发表研究成果。Gousios 等人[3]通过综合传统度量方法总结了一系列开源环境下项目成员的贡献行为,其中包括很多软件资源库,如 IRC、Wiki、Bug Database 等,但这些资源库的研究范围与国内实际情况不符。Amor 等人[4]也根据软件开发过程中的跟踪信息提出一套评估模型。此外,Treude 等人[5]提出 4 方面贡献:代码贡献、方法平均复杂度、引入的 bug、bug 修复,并收集 4 个团队、48 位开发者、为期 12 周的贡献数据,从而记录整个过程的贡献情况,以此完成验证。

2 研究内容

2.1 概述

本文从 GitHub 获取项目团队内各成员相关数据,主要从以下方面进行软件仓库挖掘,将成员在团队中的贡献具体化,可以帮助教师解决 TA 问题。首先进行理论研究:构建成员贡献行为指标模型和设计量化贡献计算方法,然后将理论成果应用于实践研究中,设计并实现 GitHub 团队项目成员贡献评估 Web 应用系统。其中,数据库采用 MySQL,后台使用 SpringMVC 框架分层解耦为持久层、业务逻辑层和控制层,最后则在前端用 Extjs 将成员贡献生成可视化展示。主要研究内容框架如图 1 所示。

本文拟将从 3 个方面展开研究阐述:成员贡献行为数据建模、成员贡献行为度量计算方法、成员贡献行为数据的获取与处理,而后将进行系统的设计与实现。这里针对各主题要点可得研究分述如下。

2.2 成员贡献行为数据建模

针对本研究,将引入如下 GitHub 开发人员贡献行为指标,设计详情参见表 1。

2.3 成员贡献行为度量计算方法

基于 Gousios 等人的研究,本文设计了一种项目团队内的成员贡献行为度量计算方法,计算公式如下所示:

$$C(d) = LOC(d) + CF(d), \tag{1}$$

其中,$C(d)$ 表示开发人员 d 的总贡献;$LOC(d)$ 表示开发人员 d 的相对净代码量;$CF(d)$ 表示开发人员 d 对表 1 中各贡献行为指标(LOC 除外)的贡献函数。

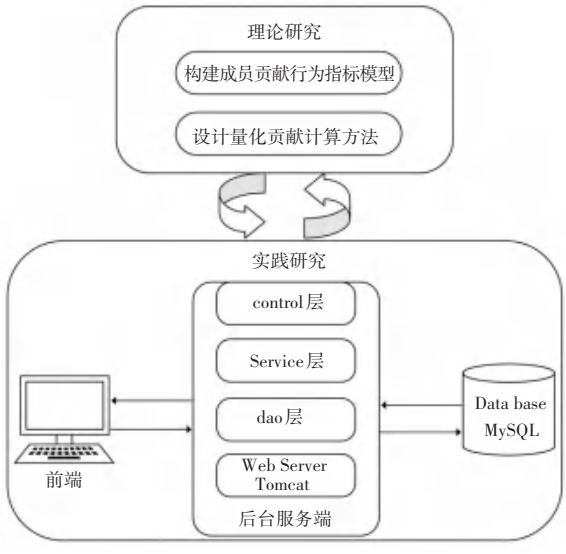

图 1 研究内容框架

Fig. 1 Research content framework

表 1 GitHub 开发人员贡献行为指标

Tab. 1 Contribution behavior indicators of GitHub developers

资源库	贡献行为指标	标识
代码及文档库[cd]	净代码量	LOC
	添加代码行	CADD
	删除代码行	CDEL
	添加文档	DADD
	删除文档	DDEL
	Commits 数	CMN
	参与项目时长	CJT
Bug[bug]	提交注释中包含 Bug 编号	CBN

进一步,研究推得 $LOC(d)$ 和 $CF(d)$ 的数学表述可分别如式(2)、式(3)所示:

$$LOC(d) = \frac{loc(d)}{\sum_{k=1}^{n} loc(k)}, \tag{2}$$

$$CF(d) = \alpha \sum_{i=1}^{n} w_i^{cd} A_i^{cd}(d) + \beta \sum_{i=1}^{n} w_i^{bug} A_i^{bug}(d). \tag{3}$$

特别地,x 表示式(3)中的 cd 和 bug。其中,$A_i^x(d)$ 表示开发人员 d 在第 x 种资源库的第 i 项贡献行为指标的贡献度;w_i^x 表示第 i 项贡献行为指标在第 x 种资源库所占的权重比例;α、β 表示第 x 种

资源库在所有资源库中所占的权重比例；n 表示对应第 x 种资源库中贡献行为指标的数目。

每种贡献行为指标的贡献度计算都是一个相对值，具体可参见表 2。

表 2　贡献行为指标的贡献度计算

Tab. 2　Calculation of contribution behavior indicators

资源库	贡献行为指标模型	贡献度计算公式
代码及文档库［cd］	添加代码行+删除代码行	$\frac{add(d)+delete(d)}{\sum_{k=1}^{n}[add(k)+delete(k)]}$
	添加文档+删除文档	$\frac{doc(d)}{totaldoc}$
	Commits 数	$\frac{commits(d)}{\sum_{k=1}^{n}commits(k)}$
	参与项目时长	$\frac{time(d)}{totaltime}$
Bug［bug］	提交注释中包含 Bug 编号	$\frac{num(d)}{totalbugfixnums}$

2.4　成员贡献行为数据的获取与处理

2.4.1　成员贡献行为数据的设计功能分析

获取学生提交到 GitHub 上的开源软件项目地址，克隆远程库。在配置好 Git 之后，根据 GitHub 输出的地址如 https://github.com/gnu-fripside/PaperManager.git，用命令 git clone 克隆对应的本地库。使用 Java ProcessBuilder 执行 git 指令，通过分析 log 日志得到所需的数据。

在分析项目团队内的各个成员对该项目的贡献之前，首先要考虑用什么来唯一确定地标识各个成员。如图 2 所示，在 GitHub 日志中有 author name、author email、committer name 和 committer email，就需要选择其中之一来标识成员的 ID。研究可知，在 Git 分布式版本管理中，作者(author)指的是实际做出修改的人，提交者(committer)指的是最后将此工作成果提交到仓库的人，当设计者为某个项目发布补丁，然后某个核心成员将该补丁并入项目时，该人即为作者，而那个核心成员就是提交者，所以要研究描述成员的贡献就要以实际做出修改的人为准则来找到贡献拥有者。另外，经过测试，日志中记录的 author name 类似于昵称，可随意更改，若成员在开发过程中更改作者名称则不能准确跟踪该成员的贡献，故而最终选择 author email 作为唯一确定标识成员的 ID。

贡献度的度量主要用到了项目的 commits 数据。commits 数据是指项目从最开始到当前获取的最新版本之间每次提交的信息，包括提交的版本号、作者的邮箱、提交时间、提交注释信息、修改的文件及内容、修改的行数统计等，下面即以 LOC 贡献为例详细阐释解析数据获取过程。

图 2　GitHub 日志中成员 ID 选择

Fig. 2　Member ID selection in GitHub logs

2.4.2　LOC 贡献设计详解

首先需要获得项目团队内成员的净代码量，换言之，即分析出项目最后一个提交版本的每行代码的作者是谁，并进行统计。本文算法的研究设计可参考代码如下。

算法：项目内可按行度量文件的路径获取算法

输入：项目所在路径及项目内文件的根目录 filePath

输出：一组可按行度量的文件路径集合

GetFiles(*filePath*)
File_List[ ] ← *Φ*
File_Root[ ] ← *filePath*
for each file in *File_Root*[ ] do
　　if file is directory and not hidden
　　then *GetFiles*(*filePath*)
　　else if file is not Binary and not hidden
　　then *File_List*[ ] ← *file*
end for

其中判断文件是否是二进制时，获取文件数据的每个字节 t，如果 $t < 32$ 且 $t!=9$ 且 $t!=10$ 且 $t!=13$ 同时成立，则该文件为二进制文件。获得的符合要求的文件目录如图 3 所示。

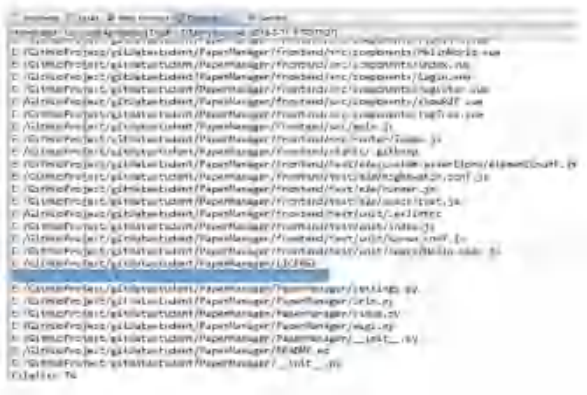

图 3 可按行度量文件目录

Fig. 3 File list measured by line

以 PaperManager 为例(项目地址为 https://github.com/gnu-fripside/PaperManager.git)。研究得到所需的项目内所有的文件目录后,就可以通过 git blame 命令逐一查看每个文件,如图 4 所示,显示内容格式为:commit ID | 作者名称 | 提交时间 | 代码位于文件中的行数|实际代码。为此将提取每行代码的 commit ID 再进行处理,上文已提到此处的作者名称不能唯一确定地标识项目内的成员,而研究提取此处的 commit ID 的目的是为了建立每行代码→commit ID→author email 的逻辑关系。

由 commit ID 关联 author email 有多种处理方式,这里可通过 git show [commit ID] --pretty=format:"%ae" - shortstat 做到更加方便快捷的处理,最后得到每行代码对应的作者邮箱,后续工作只需 HashMap 遍历计数即可得到项目中每个成员的净代码量,设计运行结果即如图 5 所示。

图 4 每个文件 git blame 的具体内容

Fig. 4 git blame content of per file

图 5 项目成员净代码量

Fig. 5 Loc of team member

3 系统设计与实现

系统首页主要内容和架构组织关系如图 6 所示。

本功能可以得到成员贡献行为度量各个指标的具体值,依次为项目团队中成员的邮箱、项目最后一个版本中成员的净代码量 LOC(以行为单位)、成员的提交数、成员参与该项目的时长(以天为单位)、成员添加和删除代码行数、成员添加和删除文档行数、提交注释中包含的 Bug 编号数。该项目功能的最终可视化效果则分别如图 7、图 8 所示,进而可以为每个指标加上权重后就能得出成员贡献排名。

图 6 系统首页

Fig. 6 System home

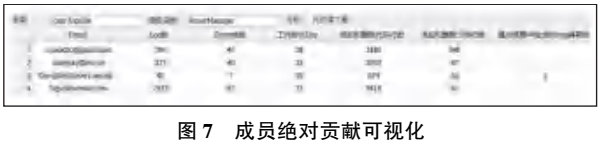

图 7 成员绝对贡献可视化

Fig. 7 Member absolute contribution visualization

图 8 成员相对贡献可视化

Fig. 8 Member relative contribution visualization

本文将参数默认设置为:不同资源库之间占比均匀且不同贡献行为指标模型之间占比均匀,由此可得项目内成员的贡献度量。同样以 PaperManager 为例(项目地址为 https://github.com/gnu-fripside/PaperManager.git)进行仿真测试,系统测试结果详见表3。由表3可见系统评分与人工评分基本为正相关,同理测试20组项目,结果证明此模型效果良好。

表3 系统测试结果

Tab. 3 The system test results

成员 ID	系统评分	人工评分(百分制)
makeztc@gmail.com	0.456 486 169	75
dongsky@live.cn	0.291 895 977	73
DongSky@users.noreply.github.com	0.664 036 637	78
lkgv@foxmail.com	0.930 438 374	85

4 结束语

本文针对 GitHub 上学生的开源团队项目构建了成员贡献行为指标模型,提出度量成员贡献的计算方法,设计并实现了基于 SpringMVC 的 GitHub 团队项目评估与监控系统的 Web 应用,利用 GitHub 软件仓库中记录的信息进行挖掘,帮助教师更好地评判学生在 GitHub 项目中的贡献。本系统还选取了20组项目,对该系统进行体验测试,结果表明本软件对项目团队内各成员的评判结果与教师人工评判基本一致,效果良好。

参考文献

[1] SACKMAN H, ERIKSON W J, GRANT E E. Exploratory experimental studies comparing online and offline programming performance[J]. Communications of the ACM, 1968, 11(1): 3-11.

[2] CARD D N, EMAM K E I. Measurement of object-oriented software development projects [Z]. Virginia: Software Productivity Consortium NFP Inc., 2001.

[3] GOUSIOS G, KALLIAMVAKOU E, SPINELLIS D. Measuring developer contribution from software repository data [C]// Proceedings of the 2008 International Working Conference on Mining Software Repositories. Leipzig, Germany: ACM, 2008: 129-132.

[4] AMOR J J, ROBLES G, GONZALEZ-BARAHONA J M. Effort estimation by characterizing developer activity[C]//Proceedings of the 2006 International Workshop on Economics Driven Software Engineering Research. Shanghai, China: ACM, 2006: 3-6.

[5] TREUDE C, FILHO F F, KULESZA U. Summarizing and measuring development activity[C]// Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering. Bergamo, Italy: ACM, 2015: 625-636.

[6] LIMA J, TREUDE C, FILHO F F, et al. Assessing developer contribution with repository mining-based metrics[C]//2015 IEEE International Conference on Software Maintenance and Evolution(ICSME). Bremen, Germany: IEEE, 2015: 536-540.

[7] LEITE L, TREUDE C, FILHO F F. UEDashboard: Awareness of unusual events in commit histories[C]// Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering. Bergamo, Italy: ACM, 2015: 978-981.

[8] ROBINSON W N, DENG Tianjie. Data mining behavioral transitions in open source repositories[C]//2015 48th Hawaii International Conference on System Sciences (HICSS). Kauai, HI, USA: IEEE, 2015: 5280-5289.

[9] YU Hao, WANG Zhongjie, CHI Xu, et al. Studying social collaboration features and patterns in service crowdsourcing [M]//SHENG Q, STROULIA E, TATA S, et al. ICSOC 2016: Service-oriented computing. Lecture Notes in Computer Science, Cham: Springer, 2016, 9936: 697-704.

[10] 徐奔. 开源软件开发人员行为特征的可视化挖掘[D]. 上海:上海交通大学, 2012.

[11] 袁霖, 王怀民, 尹刚, 等. 开源环境下开发人员行为特征挖掘与分析[J]. 计算机学报, 2010, 33(10): 1909-1918.

(上接第306页)

参考文献

[1] 刘明辉, 李玉辉, 林刚, 等. 完善厦门居家养老、医养结合、智慧养老服务体系—面对"十三五"期间养老需求[J]. 厦门科技, 2017(1): 1-5.

[2] [英]MARGARET A B. AI:人工智能的本质与未来[M]. 孙诗惠,译.北京:中国人民大学出版社, 2017.

[3] 彭露露, 杨彤蕾, 余剑. "互联网"背景下居家智慧健康养老模式的探究[J]. 价值工程, 2019, 38(4): 164-166.

[4] 闫志明, 唐夏夏, 秦旋, 等. 教育人工智能(EAI)的内涵、关键技术与应用趋势——美国《为人工智能的未来做好准备》和《国家人工智能研发战略规划》报告解析[J]. 远程教育杂志, 2017(1): 26-35.

[5] 李长远. "互联网+"在社区居家养老服务中应用的问题及对策[J]. 北京邮电大学学报(社会科学版), 2016, 18(5): 67-73.

[6] 李振, 周东岱, 刘娜, 等. 人工智能应用背景下的教育人工智能研究[J]. 现代教育技术, 2018(9): 19-25.

[7] 屈芳, 郭骅. "物联网+大数据"视阈下的智慧养老模式研究[J]. 信息资源管理学报, 2017(4): 51-57.

[8] 霍凤财,迟金,黄梓健,等.移动机器人路径规划算法综述[J]. 吉林大学学报(信息科学版),2018, 36(6): 639-647.

[9] 龚艳萍. 互联网+社区+居家养老产业发展研究—以荆门市为例的养老产业PPP项目思考[J]. 荆楚学刊, 2016, 17(1): 36-40.