

2018 年以来开源项目前沿论文最新进展

2018.11.06 方建勇

提示：采用手机 safari 微软翻译技术

1. 第 18110020[[pdf](#),[其他](#)] cse

太简单，测试？低故障风险识别方法缺陷预测的逆向看法

作者:[rainer nierdemayr](#), [tobias röhm](#), [stefan wagner](#)

摘要: 背景。测试资源通常是有限的，因此通常不可能在发布之前完全测试应用程序。为了解决资源稀缺的问题，开发团队可以应用缺陷预测来识别容易发生故障的代码区域。然而，在跨项目预测场景中，缺陷预测往往不准确。目标。我们对缺陷预测持相反的看法，旨在确定在测试时可以延迟的方法，因为它们几乎不包含任何错误，因为它们的代码是“微不足道的”。我们期望这些方法的特性可能是独立于项目的，因此我们的方法可以改进跨项目预测。方法。我们计算代码度量并应用关联规则挖掘来创建用于识别具有较低故障风险的方法的规则。我们对六个 java 开源项目进行了实证研究，以评估我们的方法，这些项目包含了方法级别的精确故障数据。结果。结果表明，逆缺陷预测能识别出约 32.44% 的项目方法具有较低的故障风险；平均而言，它们包含故障的可能性是其他方法的六倍左右。在具有更大、更多样化的训练集的跨项目预测中，确定的方法包含故障的可能性甚至要低 11 倍。结论。反向缺陷预测通过识别在测试活动中可以优先处理较少的方法，支持测试资源的有效分配，并在跨项目预测方案中很好地适用。少

2018 年 11 月 2 日提交;最初宣布 2018 年 11 月。

评论:提交给 peerj cs

类:D.2.5;D.2。8

2. 建议: 1810.09477[[pdf](#),[其他](#)] cse

莫诺雷波斯: 多元文学述评

作者:[gleison brito](#), [ricardo terra](#), [marco tulio valente](#)

摘要: 单片存储库 (monorepo) 被谷歌和脸谱等大公司以及巴贝尔和恩伯等热门开源项目所使用。本研究概述了单点的定义和特点及其好处和挑战。于是，我们对主要是灰色的文学进行了多声文学综述。我们的发现有四个方面。首先，monorepos 是包含多个相关或不相关的项目的单个存储库，它们共享相同的依赖关系。其次，集中化和标准化是一些关键特征。第三，主要优点包括简化依赖关系、跨项目更改的协调以及轻松的重构。第四，代码运行状况、代码库复杂性以及用于开发和执行的工具投资被认为是主要的挑战。少

2018 年 10 月 22 日提交;最初宣布 2018 年 10 月。

评论:发表于: 第六届巴西软件可视化、进化和维护研讨会, 第 1-8 页, 2018 年

3. 建议: 1810.0753[[pdf](#), [ps](#),[其他](#)] Cs。直流

在物联网网关上实现多容器部署

作者:[kodolui](#), [csaba kiraly](#)

摘要: 高级物联网 (iot) 应用中的严格延迟要求以及云数据中心的负载增加, 促使您转向更加分散的方法, 使物联网数据的存储和处理更接近终端设备通过部署多功能物联网网关。但是, 这些网关的资源受限性质和多样性对开发可广泛部署的应用程序构成了挑战。通过容器化 (一种轻量级虚拟化形式) 来克服这一挑战, 它为广泛的硬件体系结构和物联网网关上的操作系统无关的应用程序部署提供了支持。本文讨论了容器化的体系结构方面, 并研究了可用的容器化工具在物联网网关上下文中部署多容器的适用性。我们将容器化呈现在 agile 的背景下, 这是一个基于多容器和微服务的物联网网关开源框架, 是作为 horizon 2020 项目的一部分而开发的。我们对容器化服务进行研究, 以执行设备发现、数据管理和云集成等常见网关功能, 揭示了在使用基本映像方面为物联网网关提供容器化环境的优势用于容器内和跨容器性能优化的层次结构和图像分层。我们在本文的一组基准实验中对这些结果进行了说明。少

2018 年 10 月 4 日提交;最初宣布 2018 年 10 月。

评论:7 页, 6 个数字

4. 第: 1810.4825[[pdf](#)] [cse](#)

多伊 [10.1109/ACCESS.2018.2872669](#)

基于 github 存储库的器官移植初探

作者:[王尚文](#),[毛晓光](#),[俞觉敏](#)

文摘: 器官移植是利用与某些特定功能直接相关的代码来完成自己的程序, 它比传统的组件重用为开发人员提供了更多的便利。然而, 最近的技术面临着缺乏移植器官的挑战。因此, 我们对从 github 存储库中提取器官进行了实证研究, 以探索基于大规模数据集的移植。我们分析了 12 个具有代表性的 github 项目的统计数据, 得出的结论是: 1) 在提交中存在大量的实际器官, 并在评论中添加一个关键词;2) 本资料库中的机关主要有四种内容;3) 大约 70% 的器官易于移植。实施不同类型器官的移植策略, 我们用三种不同的编程语言 (java、python 和 c) 手动提取 30 个器官, 并利用四种测试工具 (两个用于 java, 一个用于 python, 和一个为 c)。最后, 我们将三个 java 器官移植到一个特定的平台上进行性能检查, 以验证它们是否能在新系统中正常工作。我们的策略提取的 30 个器官在单元测试中都有良好的性能, 最高合格率达到 97%, 最低的器官仍然通过 80%, 三个 java 器官在新系统中工作良好, 为主机提供了三个新的功能。结果表明, 基于开源存储库的器官移植是可行的, 为代码的重用带来了新的思路。少

2018 年 10 月 10 日提交;最初宣布 2018 年 10 月。

评论:14 页

5. 第 1810.4779[[pdf](#),[其他](#)] [Cs](#)。 铭

重定向 2 自身: 通过异地间接访问保护用户上传内容的知识产权

作者:[georgeos kon](#) 中移动于 [angelos d.Keromytis](#) , [georgeos Portokalidis](#)

摘要: 社交网络服务吸引了数百万用户, 包括个人、专业人士和公司, 每天上传大量内容, 如文字、图片和视频。内容创建者保留他们与这些网络共享的内容上的知识产权 (ip) 权利, 但是, 他们通常会含蓄地授予他们, 有时使用该内容的许可过于宽泛, 从而使服务能够在可能的情况下使用这些内容。不可取的方式。例如, facebook 声称所有用户提供的內容都有可转让的、可分许可的、免版税的全球许可。专业内容创作者和摄影师一样, 尤其受到影响。在本文中, 我们提出了一种在不影响用户功能损失的情况下将用户数据与社交网络服务脱钩的设计。我们的设计表明, 在能够在服务条款和整体环境 (例如, 不同的位置) 下托管用户生成的内容的第三方, 用户数据被排除在社交网

络服务之外, 以更好地满足用户的需求和愿望.同时, 间接架构无缝集成到社交网络服务中, 无需服务器端的任何合作, 因此用户可以像在现场托管时一样透明地访问异地数据.我们已经实现了我们的设计, 作为 chrome 浏览器的扩展, 称为 `redirect2own`, 并表明它在访问 "重定向" 内容时产生的开销可以忽略不计.我们提供的扩展作为自由软件和它的代码作为一个**开源的项目**. 少

2018 年 10 月 10 日提交;最初宣布 2018 年 10 月。

6. [建议: 1810.01791](#)[pdf,其他] cse

固定器: 挖掘用于自动程序修复的相关修复模式

作者:[anil koyuncu](#), [kui liu](#), [tegawendéf. bissyardé](#), [dunsun kim](#), [jacques khalin](#), [martin monperrus](#), [yves le traon](#)

摘要: 代码理解在软件维护中至关重要。为了提供支持维护任务的工具和方法, 研究人员研究了与如何以抽象形式描述软件代码有关的各种研究领域。到目前为止, 关于变化模式挖掘、代码克隆检测或语义补丁推理的研究主要采用基于文本、令牌和树表示作为计算代码片段相似性的基础。尽管一般来说, 现有的技术形成了 "类似" 代码的集群, 但我们在修补程序挖掘方面的经验表明, 由这些技术形成的修补程序集群通常不包含可与错误修复模式相关的可解释的语义。本文提出了一种基于迭代、三重聚类策略的新的自动化方法来挖掘与语义相关的修复模式。我们的技术 `fixminer` 为每个群集圆利用不同的树表示: 抽象语法树、编辑操作树和代码上下文树。我们对从**开源项目**中收集的数千个软件补丁进行了修复。初步结果表明, 我们能够挖掘出准确的模式, 有效地利用 `ast` 差异树中的变化信息。最终, `fixminer` 会产生一些模式, 这些模式可以与关联的修补程序所处理的 `bug` 的语义相关联。我们进一步利用挖掘的模式来实现自动化的程序修复管道, 我们能够正确地修复来自 `defect4j` 基准的 25 个错误。除了这种定量性能之外, 我们还表明, 挖掘的修复模式具有足够的相关性, 可以生成具有很高正确性概率的补丁: 固定人员生成的可信补丁的 80% 是正确的, 而最接近的相关工作, 即 `hdfix` 和 `simfix`, 分别实现 26% 和 70% 的正确性。少

2018 年 10 月 3 日提交;最初宣布 2018 年 10 月。

评论:31 页, 11 位数字

7. [建议: 1810.01662](#)[pdf,其他] Cs. 铬

面向 `javacard` 的低级加密原语

作者:[vasilios mavroudis](#), [petr svenda](#)

摘要: `javacard` 是一个多应用程序安全平台, 部署到超过 200 亿张智能卡, 用于从安全支付到电信的各种应用程序。虽然该平台是已建立的商业用例 (例如电信网络中的 `sim` 卡) 的热门选择, 但它在以下应用场景中的采用率明显较低: 1) 需要最近标准化的加密算法, 2) 研究项目, 以及 3) **开源** 计划。我们将此归因于对低级加密基元 (例如椭圆曲线操作) 的受限访问, 以及缺乏基本数据类型 (例如整数)。虽然底层硬件具有这些功能, 但 `javacard api` 不提供对相应功能的调用。到目前为止, 唯一可用的解决方法是制造商特定的专有 `api`, 这些 `api` 附带了非常严格的保密协议。在本文中, 我们引入了一种从高级操作中有效地派生基本数据类型和低级加密基元的方法。我们的技术是资源受限平台的理想选择, 并可充分利用底层硬件, 同时占用较小的内存。我们还介绍了 `jcmathlib`, 据我们所知, `jcmathlib` 是 `javacard` 中第一个不依赖于专有 `api` 的低级加密操作的通用库。在没有任何公开限制的情况下, `jcmathlib` 支持**开放式**代码共享、研究原型发布以及公共和第三方代码审核。少

2018 年 10 月 3 日提交;最初宣布 2018 年 10 月。

8. [第: 1810.1248](#)[\[pdf,其他\]](#) Cs. Sd

轻量级音乐纹理传输系统

作者:[彭秀坦](#)、[陈丽](#)、[蔡志](#)、[石法强](#)、[刘一丹](#)、[李建新](#)

摘要: 对图像和文本转换问题的深度学习研究引起了人们的极大关注。然而,目前利用神经网络进行音乐特征传输的方法还远远没有实际应用。本文提出了一种新的音乐纹理传输系统,并将其作为开源项目发布。其核心算法由表示声音为纹理谱的转换器、相应的重构器和前馈传输网络组成。我们从多个角度对该系统进行了评价,实验结果表明,该系统在音效和计算性能方面都取得了令人信服的效果。少

2018 年 9 月 27 日提交;最初宣布 2018 年 10 月。

评论:12 页

9. [第 1810.00101](#)[\[pdf,其他\]](#) cse

java 项目中的内存和资源泄漏缺陷及其修复

作者:[mohamrerreza ghanavati](#), [diego costa](#), [janos seboek](#), [david lo](#), [artur andrzejak](#)

摘要: 尽管进行了大量的软件工程工作和编程语言支持,但资源和内存泄漏仍然是一个棘手的问题,即使在内存管理的语言(如 java)中也是如此。了解泄漏诱发缺陷的特性、泄漏的明显性以及如何修复这些缺陷是设计更好的方法以避免、诊断和修复与泄漏相关的缺陷的必要先决条件。我们对 10 个大型开源 java 项目的 452 个问题进行了详细的实证研究。该研究提出了泄漏类型、造成泄漏类型的缺陷和修复措施的分类。我们从几个方面研究每个分类中的分布以及它们之间的关系。我们发现,手动代码检测和手动运行时检测仍然是泄漏检测的主要方法。我们发现,大多数错误都是在无错误的执行路径上表现出来的,与非泄漏缺陷相比,开发人员在更短的时间内修复泄漏缺陷。我们还在修复修补程序中确定了 13 个重复出现的代码转换。根据我们的发现,我们对开发人员如何避免、检测、隔离和修复与泄漏相关的错误产生了各种影响。少

2018 年 9 月 28 日提交;最初宣布 2018 年 10 月。

10. [第 1809.10265](#)[\[pdf\]](#) cse

通过图文图进行软件发布分析的基础结构

作者:[felipe curty](#), [troy kohwalter](#), [vanessa braganholo](#), [leonardo murta](#)

摘要: 如今,通过持续交付流程快速发展和交付软件是一种竞争优势,也是根据客户需求的频繁变化来更新软件的一种方式。然而,软件发布周期越快,跟踪软件演变的难度就越大。在本文中,我们提出了中继简单,这是一个旨在通过生成和报告项目发布来源来支持使用连续交付的项目的工具。resueasy 生成的来源允许对软件演变进行图形可视化,并支持查询以发现隐式信息,例如每个版本的实现功能和相关开发人员。本文还展示了对“请信”在行动中的初步评价,利用我们的工具收集的来源生成了开源项目的变更日志。少

2018 年 9 月 26 日提交;最初宣布 2018 年 9 月。

评论:2018 年第六届巴西软件可视化、进化和维护研讨会 (vem) 接受

11. [第 1809.09851](#)[\[pdf, ps,其他\]](#) cs. ms

fdbb: 流体动力学建筑块

作者:[matthias möller](#), [andrzej jesechke](#)

摘要: 高性能计算平台变得越来越异构, 这使得研究人员和科学软件开发人员很难跟上硬件市场的快速变化。本文提出了**基于开源的 fdbb** (流体动力学构建块) 项目, 为异构系统的流体动力学应用提供了一条便利。它由一个低级 api 组成, 它为许多不同的线性代数后端提供了一个统一的接口, 并提供了一个轻量级和可扩展的高级表达式模板库, 它提供了很大程度上可定制的流体动力学构建块, 例如主要变量和次要变量之间的变换, 以及黎曼不变性、状态方程、无粘性通量及其通量-雅各布人的表达式。对所开发方法的性能进行了综合微基准和小型应用的评估。少

2018 年 9 月 26 日提交;最初宣布 2018 年 9 月。

12. 第 [xiv:18.099.4774](#)[pdf,其他] Cs. 铭

fideliuS: 保护用户机密不受被破坏的浏览器的影响

作者:[saba Eskandarian](#), [jonathan cogan](#), [sawyer birnbaum](#), [peh changwei](#) brandon, [dillon franke](#), [forest fraser](#), [gaspar garcia jr.](#), [eric gong](#), [hung t. nguyen](#), [taresh k. sethi](#), [vishal subbiah](#), [michael backes](#), [giancarlo pellegrino](#), [dan boneh](#)

摘要: 用户定期在浏览器窗口中输入敏感数据, 如密码、信用卡号或税务信息。虽然现代浏览器提供了强大的客户端隐私措施来保护这些数据, 但这些防御措施都不能阻止被恶意软件破坏的浏览器窃取它。在这项工作中, 我们介绍了 **fideliuS**, 这是一种新的体系结构, 它使用集成到浏览器中的受信任硬件飞地, 以便在 web 浏览会话期间保护用户机密, 即使整个底层浏览器和操作系统完全由恶意浏览器控制攻击。**fideliuS** 解决了在完全恶意环境中为浏览器提供保护、为表单数据的完整性和隐私提供支持、javascript 执行、XMLHttpRequests 和受保护的 web 存储, 同时最大限度地减少了 tcb。此外, 飞地与浏览器、键盘和显示器之间的交互都需要新的协议, 每个协议都有自己的安全考虑。最后, **fideliuS** 考虑了 ui 因素, 以确保开发人员和用户的界面一致而简单。作为该项目的一部分, 我们开发了第一个**开源**系统, 该系统提供了从输入和输出外围设备到硬件飞地的受信任路径, 而不依赖于其他虚拟机管理程序安全假设。这些组成部分可能具有独立的意义, 对今后的项目有用。我们实现并评估 **fideliuS** 以测量其性能开销, 发现 **fideliuS** 对受保护页面的页面加载和用户交互施加了可接受的开销, 并且对不使用其增强安全性的页面和页面组件没有影响特征。少

2018 年 9 月 17 日提交;v1 于 2018 年 9 月 13 日提交;最初宣布 2018 年 9 月。

13. 第 [xiv:1809.04424](#)[pdf, ps,其他] cs. ms

调整计算持久性同源包的性能

作者:[alan hylton](#), [gregory Henselman-Petrusek](#), [janche sang](#), [robert short](#)

摘要: 近年来, 持久的同源性已成为一种有吸引力的数据分析方法。它从点云数据中捕获拓扑特征, 如连接的组件、孔和空隙, 并总结这些功能在过滤序列中出现和消失的方式。在这个**项目**中, 我们专注于提高 **eirene** 的性能, 这是一个持久同源的**计算包**。**艾琳**是一个 5000 行**开源软件库**, 采用动态编程语言 **julia** 实现。我们使用 **julia** 分析工具来识别性能瓶颈, 并开发新的方法来管理这些瓶颈, 包括在多核核心硬件上并行化一些耗时的功能。实证结果表明, 性能有了很大的提高。少

2018 年 9 月 8 日提交;最初宣布 2018 年 9 月。

评论:29 页, 10 个数字

14. 第 [xiv:1809.04177](#)[pdf,其他] Cs. Hc

在线课程中的点击流日志的时间序列分析

作者:[yohanjo](#), [keith maki](#), [gaurav tomar](#)

摘要: 由于大规模开放在线课程 (mooc) 的普及程度迅速提高, 对学生学习的可扩展自动化支持技术的需求日益增加。近年来, 传统教育资源向网络环境转移已成为一个越来越重要的问题。为了使学习科学理论得到应用, 教育工作者需要一种方法来识别学生的学习行为, 这些行为有助于学习结果, 并利用这些行为来设计和为学生提供个性化的干预支持。点击日志是学生学习行为的重要信息来源, 但目前的文献对点击日志中如何表达这些行为的理解有限。在本项目中, 我们利用学生行为的时间动态, 通过图形建模方法进行行为建模, 并通过反复的神经网络方法进行性能预测, 以便首先识别学生行为, 然后使用它们来预测他们在课程中的最终结果。我们的实验表明, 长期短期记忆 (lstm) 模型能够在序列中学习长期依赖关系, 并优于预测任务中的其他强基线。此外, 当与从图形模型行为建模中获得的结果一起使用时, 可以成功地将这些单击日志分析的顺序方法导入到其他课程中。少

2018 年 9 月 11 日提交;最初宣布 2018 年 9 月。

评论:技术报告

15. 第 [xiv:1809.04041](#)[pdf,其他] cse

在 github 中识别未维护的项目

作者:[jailton coelho](#), [marco tulio valente](#), [lu 西亚 a l.silva](#), [emad shihab](#)

摘要: 背景:开源 软件在现代软件开发中的重要性日益增加。然而, 人们对这类项目的可持续性也越来越感到关切, 这些项目通常由少数开发商管理, 经常担任志愿者。目标:在本文中, 我们提出了一种方法来识别没有积极维护的 github 项目。我们的目标是提醒用户使用这些项目的风险, 并可能激励其他开发人员承担项目的维护。方法: 我们根据一组有关项目活动的功能 (提交、叉、问题等), 训练机器学习模型, 以识别不维护或维护稀少的项目。我们与 129 个 github 项目的主要开发人员以最佳性能对模型进行了经验验证。结果: 基于实际开源 开发者的反馈, 提出的机器学习方法的精度为 80%; 和 96% 的召回。我们还表明, 我们的方法可以用来评估项目得不到维护的风险。结论: 本文提出的模型可供 开源用户和开发人员用于识别不再积极维护的 github 项目。少

2018 年 9 月 11 日提交;最初宣布 2018 年 9 月。

评论:2018 年在第十二届经验软件工程与测量国际研讨会 (esem) 上接受, 10 页

16. 第 [xiv:1809.03725](#)[pdf,其他] cse

开源开发人员社区的多样性、生产力和增长

作者:[清业江](#), [杨春利](#), [约瑟夫·g·戴维斯](#), [阿布莱特·佐马亚](#)

摘要: 开源 开发模型已经成为传统的内部源软件开发模式的范式转变, 取得了许多成功。传统上,开源 项目的主要特点是他们的个人志愿者开发人员。特别是在许多组织的参与下, 这种传统发生了重大变化。但是, 在开源开发人员社区如何演变方面存在知识差距。在本文中, 我们对开源 开发人员社区提出了一些看法。特别是, 我们分析了 20 个知名开源 项目的 git 存储库, 总共有 300 多万项提交活动。利用斯皮尔曼的等级相关系数、多样性指数和贡珀尔齐逻辑曲线分别对生产率、多样性和生长进行了分析。我们发现 (a) 斯皮尔曼的等级相关系数之间的积极贡献者和承诺活动揭示了开发人员社区规模的变化如何影响社区的生产力;(二)开源 发展商社区的多元化指数显示该社区的架构; 及(c) 开源开发者社区的增长可以用不同的生长曲线阶段来描述, 就像在许多有机物质中一样。少

2018 年 9 月 11 日提交;最初宣布 2018 年 9 月。

评论:10 页, 6 个数字

17. 文件: 1809.03125[[pdf](#), [ps](#),其他] Cs. 红外

多伊 [10.18122/cs_facpubs/147/boisestate](#)

推荐系统实验的 lppy 包: 下一代工具和从 lenskit 项目中吸取的经验教训

作者:michael d. ekstrand

摘要: 自 2010 年以来, 我们一直在构建和维护 lenskit, 这是一个用于构建、研究和学习推荐系统的开源工具包。我们已成功地将该软件用于广泛的推荐系统实验, 以支持传统课堂和在线设置中的教育, 并作为电影和书籍中面向用户的推荐服务的算法后端。这一经历, 加上社区反馈, 在 lenskit 的设计和选择方面出现了一些挑战。为了应对这些挑战, 我们正在开发一套新的工具, 利用 pydata 堆栈, 使我们能够通过 lenskit 提供的各种研究实验和教育经验, 以及新的实验结构, 现有的代码使困难。其结果是一组研究工具, 应显著提高研究速度, 并提供更平滑的集成与其他软件, 如 keras, 同时保持与 lenskit 实验相同的重现性水平。在本文中, 我们回顾了 lenskit 项目, 特别是我们将其用于离线评估实验的经验, 并描述了下一代 lkpy 工具, 以便能够使用灵活、开放的离线评估和实验-结束设计和经过良好测试的评估原语。少

2018 年 9 月 10 日提交;最初宣布 2018 年 9 月。

评论:在 2018 年推荐系统离线评估回收研讨会上发表

18. 第 09iv:1800.005554[[pdf](#),其他] Cs. 直流

yac:bft 区块链一致性算法

作者:fedor muratov, andrei lebedev, nikolai iushkevich, bulat nasrulin, makoto takemiya

摘要: 在异步接收事件并受到拜占庭故障影响的分散系统中, 共识是许多实际应用中的常见问题。分散系统的进步, 如分布式分类账 (即区块链) 技术, 只是增加了在分散系统中寻找性能和安全的解决方案以达成状态机复制共识的重要性。yac 是一种实用的分散共识算法, 它解决了经典拜占庭容错共识算法中出现的消息传递效率低下和领导者强的问题。该算法是开源的, 目前用于为超分类账 iroha 区块链项目提供拜占庭容错共识。我们提供了安全性和活性的证明, 以及经验结果表明, 我们的算法可以扩展到几十个验证对等方。少

2018 年 9 月 3 日提交;最初宣布 2018 年 9 月。

19. 第 xiv:18009.00143[[pdf](#),其他] cse

多伊 [10.1016/j.jss.2018.08.061](#)

持续集成环境中的测试优先级

作者:alireza haghatakhah, mika mäntylä, markku oivo, pasi kuvaja

摘要: 研究是为了更早地捕获回归错误, 使开发人员能够更频繁、更持续地集成和验证他们的更改。为了实现这一目标, 我们调查了六个开源项目, 每个项目都包括在很长一段时间内的几个构建。结果表明..。更多

2018 年 9 月 1 日提交;最初宣布 2018 年 9 月。

评论:系统与软件杂志 (2018)

日记本参考:系统与软件杂志 (2018)

20. 第 1808. 10703[[pdf](#)] 反渗透委员会

python 机器人: 机器人算法的 python 代码集合

作者:[atsushi sakai](#), [daniel ingram](#), [joseph dinius](#), [karan chawla](#), [安东南·拉芬](#), [亚历克西斯·帕克斯](#)

文摘: 本文介绍了一个开源 软件 (oss) 项目: pythonrobotics。这是在 python 编程语言中实现的机器人算法的集合。该项目的重点是自主导航, 目标是让机器人初学者了解每个算法背后的基本思想。在本项目中, 选择了实用且在学术界和产业界广泛使用的算法。每个示例代码都是用 python3 编写的, 并且仅依赖于一些标准模块的可读性和易用性。它包括直观的动画, 以了解模拟的行为。少

2018 年 9 月 7 日提交;v1 于 2018 年 8 月 31 日提交;最初宣布 2018 年 8 月。

评论:8 页, 6 个数字, 这是代码存储库:

<https://github.com/AtsushiSakai/PythonRobotics>

21. [第 1808. 10033\[pdf,其他\]](#) [cse](#)

源代码相似度指标在软件缺陷预测中的应用

作者:[阿赫梅特·奥库坦](#)

摘要: 近年来, 缺陷预测在经验软件工程领域受到了广泛的关注。在维护阶段之前预测软件缺陷不仅对降低维护成本, 而且对提高软件产品的整体质量都非常重要。到目前为止, 有不同类型的产品、流程和基于开发人员的软件指标被提出来衡量软件系统的缺陷。本文建议使用一套新的软件度量方法, 该度量方法基于软件项目中检测到的源代码文件之间的相似性。为了寻找软件系统不同文件之间的源代码相似性, 采用了剽窃和克隆检测技术。考虑到每个文件与项目中有缺陷和无缺陷文件的总体相似性, 为每个文件计算了两个简单的相似度指标。使用这些相似性指标, 我们可以预测特定文件是否有缺陷。我们在 10 个开源数据集上的实验表明, 根据检测到的相似度, 与现有的静态代码指标相比, 建议的指标在以下领域可以获得更好的性能。曲线 (auc)。少

2018 年 8 月 29 日提交;最初宣布 2018 年 8 月。

评论:一种利用源代码相似度度量进行软件缺陷预测的新方法

22. [特别报告: 1808. 07937\[pdf,其他\]](#) [cs.PL](#)

使用合同在 erlang 中进行运行时验证

作者:[lars-奥克·弗莱德](#), [julio mariño](#), [sergio pérez](#), [salvador tamarit](#)

摘要: 在程序的生存期内, 程序会遭受一些变化, 这些更改旨在改进或增加其功能的某些部分。但是, 这些修改通常也会导致影响已工作的代码的错误。有几种方法和工具可以帮助发现和发现这些错误的来源。但是, 大多数这些错误可以通过使用程序员在编写代码时所拥有的一些知识来避免。这就是按合同设计方法的思想, 用户可以在运行时定义可在运行时检查的协定。在本文中, 我们将这种方法的原理应用于 erlang, 从而使这种语言中的运行时验证系统成为可能。我们定义了两类类型的合同。其中之一可以在任何 erlang 程序中使用, 而第二种类型仅用于并发程序。我们提供了这两类合同实施的详细信息。此外, 我们还对每份合同进行了广泛的解释, 并举例说明了合同的使用情况。本文提出的所有思想都已在基于契约的运行时验证系统中实现, 该系统名为 edbc。它的源代码可在 github 作为一个开源和免费的项目。少

2018 年 8 月 27 日提交;v1 于 2018 年 8 月 23 日提交;最初宣布 2018 年 8 月。

评论:19 页, 可在 2018 年 wflp 中提交

23. [xiv:1808. 06489\[pdf\]](#) [cs. cy](#)

软件公司是协作还是竞争？社区发起的开放源码软件项目中的合作模式

作者: [anh nguyen-duc](#), [daniela s.cruzes](#), [snarby terje](#), [pekka abrahamsson](#)

摘要: [背景]越来越多的商业公司正在参与开放源码软件项目，以降低开发成本和提高技术创新能力。当与其他追求价值是利益冲突的公司合作时，公司的行为可能不合作，导致对共同目标的有害影响。[瞄准]本研究探讨了软件公司如何在开放源码软件项目中进行协作和竞争。[方法]我们对三个开放源码软件项目采用了混合研究方法。[结果]我们发现，参与社区发起的开放源码软件项目的商业公司跨越组织边界以各种方式进行协作。虽然大多数公司的贡献很少，但少数公司非常活跃，占捐款的很大比例。我们提出了一个概念模型来解释开放源码软件项目中软件公司之间的合作。该模型表明，在坚定的守门人的基础上，可以同时管理合作的两个方面。[结论]企业需要实施其合作战略，以最大限度地参与开放源码软件项目中获得价值。少

2018 年 8 月 16 日提交;最初宣布 2018 年 8 月。

评论:arxiv 管理说明: 文本与 [arxiv:1711.07049](#) 重叠

期刊参考:电子信息软件工程学报, 2018

24. 第: 1808. 06662[[pdf](#),其他] cse

利用链接预测技术对建筑气味的预测

作者: [j. andrés díaz-pace](#), [antonela tommasel](#), [daniela godoy](#)

摘要: 软件系统自然演变，这种演变通常会带来导致系统退化的设计问题。建筑气味是此类问题的典型症状，其中一些气味与模块之间不希望的依赖关系有关。早期检测这些气味对开发人员来说非常重要，因为他们可以提前计划维护或重构工作，从而防止系统退化。识别体系结构气味的现有工具可以检测源代码中存在的气味。这意味着它们不需要的依赖项已经创建。在本工作中，我们探索了一种前瞻性的方法，该方法能够推断可能存在的模块依赖关系组，这些依赖关系可以在未来的系统版本中预测体系结构气味。我们的方法将当前模块结构与以前版本的信息以及信息结合起来，应用链接预测技术（来自社交网络分析领域）。特别是，我们关注与依赖相关的气味，如循环依赖和 hub 式依赖，它们与链接预测模型非常吻合。对两个开源项目的初步评估表明，在某些考虑下，我们的方法的预测是令人满意的。此外，该方法还可以扩展到其他类型的基于依赖的气味或指标。少

2018 年 8 月 20 日提交;最初宣布 2018 年 8 月。

评论:可在 2018 年 [scam](#) 上发布

25. 第: 1808. 06359[[pdf](#),其他] cse

利用需求和源代码之间的历史关联来识别受影响的类

作者: [davedde falessi](#), [justinroll](#), [jin guo](#), [jane Cleland-Huang](#)

摘要: 随着软件系统中引入和实现新的要求，开发人员必须确定需要更改的源代码类集。因此，过去的工作重点是预测受需求影响的类集。在本文中，我们引入和评估了一种新的信息类型，基于这样的直觉，即与特定类的历史变化相关的一组需求可能会表现出与影响新需求的语义相似性。类。这种对需求集 ($r2rs$) 系列度量的新要求捕获了新需求与以前与类关联的现有需求集之间的语义相似性。本文的目的是介绍和评估 $r2rs$ 度量在预测受需求影响的类集合方面的有用性。我们通过结合六种自然语言处理技术来测量文本之间的语义相似度（如 vsm ）和三个分布分数来计算总体相似性（例如，相似性分数之间的平均值），考虑 18 个不同的 $r2rs$ 指标。我们评估 $r2rs$ 是否可用

于组合预测受影响的类, 以及是否与基于更改的时间位置、代码的直接相似性、复杂性度量和代码气味的其他四个度量系列进行预测。我们的评估具有五个分类器和 78 个版本属于四个大型**开源项目**, 这导致超过 700, 000 个候选受影响的类。实验结果表明, 利用 r2rs 信息可以在各种分类器和**项目**中平均提高 60% 以上的预测受影响类的准确性。少

2018 年 8 月 20 日提交;最初宣布 2018 年 8 月。

26. 第: 1808. 06292[[pdf](#)] **cs.PL**

摇滚底部, 世界, 天空: 卡特罗比, 一个极其大规模和长期的视觉编码项目完全依靠智能手机

作者:[kirshan kumar luhana](#), [matthias mueller](#), [christian schindler](#), [wolfgang slani](#), [bernadette spieler](#)

摘要: 世界各地 7 亿青少年中的大多数已经拥有自己的智能手机, 但相对较少的青少年可以使用 pc、笔记本电脑、olpc、Chromebooks 或平板电脑。免费的**开源非营利项目** catrobat 允许用户只使用智能手机创建和发布自己的应用程序。catrobat 于 2010 年启动, 自 2014 年以来推出了我们的免费应用的第一个公共版本, 自 2018 年 7 月起发布了 47 个主要编码应用程序版本, 目前有来自 180 个国家/地区的 70 多万用户, 有 50 多种语言提供服务, 迄今已开发了近 1, 000 个来自世界各地 ("世界") 的志愿者。catrobat 受到 scratch 的强烈启发, 确实允许导入大多数 scratch **项目**, 因此, 截至 2018 年 7 月, 我们的用户手机上可以访问 3, 000 多万个**项目**。我们的应用程序非常直观 ("底部"), 有许多辅助功能设置, 例如, 对于有视觉或认知障碍的孩子, 并且有大量的建构主义教程和课程在许多语言。我们还创建了大量的扩展, 例如, 为各种教育机器人, 包括乐高头脑风暴和飞行鹦鹉四轮车 ("天空"), 以及控制任意外部设备通过 arduino 或 raspberry pi 板, 上升到同温层, 甚至超越行星际空间 ("天空")。目前正在开发一种 turtletitch 扩展, 允许为自己的刺绣图案编码。其中, 卡特罗比强烈关注包括女性青少年。虽然正在开发专门的学校版本, 但我们的应用程序主要用于教室外、任何地方, 特别是户外 ("底部"、"世界")。catrobat 是我们的用户通过 google play 等各种应用商店和 youtube 等社交媒体渠道以及我们在 code. org 上的存在发现的。少

2018 年 8 月 19 日提交;最初宣布 2018 年 8 月。

评论:2018 年建构主义

27. 第 **xiv: 1808. 04673**[[pdf](#),[其他](#)] **cse**

从代码存储库问题和错误报告中挖掘开源项目和库的威胁情报

作者:[lorenzo neil](#), [sudip mittal](#), [anupam joshi](#)

摘要: **开源项目**和库被用于软件开发, 同时也存在多个安全漏洞。这种对第三方生态系统的使用为开发中的产品创造了一种新的攻击面。智能攻击者可以通过利用链接**项目**和库中存在的漏洞之一来攻击产品。在本文中, 我们从公共代码存储库中报告的错误和问题中挖掘有关**开源项目**和库的威胁情报。我们还跟踪客户端计算机上已安装软件的库和**项目**依赖关系。我们在安全知识图中表示和存储此威胁情报以及软件依赖关系。然后, 如果发现有关链接库和项目的任何威胁情报, 安全分析师和开发人员可以查询和接收知识图中的警报, 这些信息在其产品中使用。少

2018 年 8 月 9 日提交;最初宣布 2018 年 8 月。

28. 第 **xiv: 1808. 03380**[[pdf](#),[其他](#)] **Cs**. 直流

对流行的加密货币中的数据传输和存储技术进行了调查, 并提出了改进建议

作者:阳光卡特库里

摘要: 本文主要研究三个相对流行的加密货币--以太、纳米和 *iota*--的八卦网络的运作。我们将研究诸如新节点加入网络时自动发现对等方、节点的带宽使用情况、消息传递协议和存储架构以及共享分类帐的优化等主题。我们认为, 在有关区块链和加密货币的工作中, 这个主题经常被忽略。由于这些项目涉及相当直接的财政问题, 漏洞和低效率比常规开源项目中的漏洞具有更高的意义。除了比特币 (一个已经存在了近 10 年的网络) 之外, 除了源代码存储库中分散和稀疏的页面之外, 没有任何其他项目对其操作细节有实质性的文档。这里描述的几乎所有内容都是通过研究这些项目的参考实现的源代码来提取的。我们评估使用可扩展的 bloom 查找表和石墨烯协议来减少块传播时间和某些消息的带宽使用。我们执行逼真的模拟, 显示出显著的改进。我们提供了一个完整的石墨烯在格思, *ethereum* 的主要节点软件, 并测试了这一实现的主要 *ethereum* 区块链。我们还为公开可见节点对所选的加密货币网络进行爬网, 并提供这些节点的自治系统级细分, 最终目标是估计执行 *bgp* 劫持等攻击的难易程度及其影响。为在 *geth* 中实现石墨烯、执行各种模拟和其他杂项任务而编写的代码已在 <https://github.com/sunfinite/masters-thesis> 上传到 github。少

2018 年 8 月 9 日提交;最初宣布 2018 年 8 月。

29. 建议: 1808. 03292[[pdf](#),其他] Cs. 马

nl4py:用于并行网络徽标工作区的 python 中基于代理的建模

作者:[chathika gunaratne](#), [ivan garibay](#)

摘要: *nl4py* 是 python 的 *netlogo* 控制器软件, 用于快速、并行地执行 *netlogo* 模型。*nl4py* 通过 python 提供无头 (无图形用户界面) 和 *gui netlogo* 工作区控制。在 python 包索引上的开源计算和机器学习库越来越多的基础上, 对通过 python 快速、并行地执行基于代理的模型的需求越来越大。*netlogo* 是大多数基于代理的建模驱动研究项目的首选语言, 它需要集成到 python 中, 以便研究人员希望使用这些库对基于代理的模型输出进行统计分析。不幸的是, 直到最近推出的 *pynetlogo*, 现在 *nl4py*, 这样的控制器是不可用的。本文详细介绍了 *nl4pi* 的使用情况, 并解释了其客户端-服务器软件体系结构, 强调了 *pynetlogo* 的体系结构差异。然后通过 *nl4py* 对狼羊捕食模型进行全球灵敏度分析和参数校准的逐步演示。最后, *NL4Py*'s 的性能以 *pynetlogo* 及其与 *IPyParallel* 的组合为基准, 并证明在这两种配置上都能显著节省执行时间。少

2018 年 8 月 20 日提交;v1 于 2018 年 8 月 9 日提交;最初宣布 2018 年 8 月。

30. 第 xiv: 1808. 01729[[pdf](#),其他] cse

可执行触发器操作注释

作者:[penyu ie](#), [rai rai](#), [junyi jsy li](#), [sarfraz khurshid](#), [raymond j . mooney](#), [milosgligoric](#)

摘要: 自然语言元素 (例如, *todo* 注释) 经常用于在开发人员之间进行通信, 并在代码存储库 (触发器) 中包含特定条件时描述需要执行的任务 (操作)。随着项目的发展、开发过程的变化和开发团队的重组, 这些评论由于其非正式性质, 往往变得无关紧要或被遗忘。我们提出了第一种称为 *trigit* 的技术, 将触发器操作到执行注释指定为可执行语句。因此, 当触发器计算为 *true* 时, 将自动执行操作。规范是用宿主语言 (如 *java*) 编写的, 并作为生成过程的一部分进行评估。这些触发器被指定为抽象语法树上的查询语句和生成配置脚本的抽象表示形式, 并且这些操作被指定为代码转换步骤。我们为 *java* 编程语言实现了 *trigit*, 并从 8 个流行的开源项目中迁移了 20 个现有的触发器操

作注释。我们根据可执行注释中的令牌数量和生成过程中引入的时间开销来评估使用 triit 的成本。少

2018 年 8 月 6 日提交;最初宣布 2018 年 8 月。

31. 第 1807. 10953[[pdf](#), [ps](#),其他] cse

多伊 [10.114/3278186.3278190](#)

基于焦点方法的面向目标的突变测试

作者:[sten vercammen](#), [mohammad ghafari](#), [serge demeyer](#), [markus borg](#)

摘要: 突变测试是评估测试套件故障检测能力的最先进技术。不幸的是, 突变测试消耗了大量的计算资源, 因为它为每个注入的突变体运行整个测试套件。本文在方法层面探讨了细粒度可追溯性链接 (称为焦点方法), 以减少突变测试的执行时间, 并验证每个方法的测试用例的质量, 而不是通常验证的整体测试套件质量。验证我们在开源 [apache 蚂蚁项目](#)上的方法显示, 在质量分数为 80% 的焦点方法中的突变体的速度加快了 5375 倍。少

2018 年 10 月 9 日提交;v1 于 2018 年 7 月 28 日提交;最初宣布 2018 年 7 月。

评论:[a-test](#) 2018

32. 第 [xiv:1807.09696](#)[[pdf](#), [ps](#),其他] cse

快速灵活的 io: 高性能设备存储结构的组合方法

作者:[丹尼尔·沃丁顿](#)

摘要: 多年来, 构建存储系统一直是系统专家的领域。它们很复杂, 难以实施。需要格外小心, 以确保性能和操作正确性的必要保证。此外, 由于基于内核的设计所施加的限制, 许多传统的实现都将软件灵活换为性能。它们的实现仅限于编译的语言 (如 c 和汇编器), 重用往往很困难或受到限制。然而, 存储系统在隐式上非常适合软件重用和组合软件的构建。有许多逻辑函数, 如块分配、缓存、分区、元数据管理等, 在大多数存储变体中都很常见。在本文中, 我们介绍了 compositional, 这是一个[开源项目](#), 它考虑了组合设计和重用以及高性能的需要, 作为一流的考虑。少

2018 年 7 月 25 日提交;最初宣布 2018 年 7 月。

评论:[7 页, 4 个数字](#)

33. 第 1807. 09224[[pdf](#),其他] Cs。哦

FluidDyn: 用于流体动力学研究和教学的 python 开源框架

作者:[pierre augier](#), [ashwin vishnu mohan](#), [cyrille bonamy](#)

摘要: FluidDyn 是一个在流体动力学社区中培养[开放科学和开源](#)的项目。它被认为是一个研究项目, 引导[开源动态](#)、方法和工具来做科学。我们提出了一套 python 包形成一个框架, 以研究流体动力学与不同的方法, 特别是实验室实验 (包流体实验室), 模拟 (包流体, 流体和流体泡沫) 和数据处理 (包流体)。在本文中, 我们概述了项目的专业包, 然后重点介绍了称为 fluiddyn 的基本包, 其中包含专门包中使用的通用代码。在 [mohan](#) 等人 (2018a. b) 的两份配套论文中更详细地介绍了[秀式流体和流体](#)。通过 FluidDyn 项目, 我们证明了专门的科学代码可以用[开源社区](#)的方法和良好做法来编写。在 [bitbb](#) (<https://bitbucket.org/fluiddyn/>) 中提供了[汞存储库](#)。所有代码都使用 [spinx](#) 和 [read](#) 文档进行记录, 并通过在 [bit](#) 存储桶、管道和 [travis](#) 上连续集成进行测试。为了提高重用潜力, 利用 python 的简单面向对象编程模型, 代码尽可能模块化。

所有代码也都是高效编写的, 使用 c++、cython 和 pythran 来提高关键函数的性能。
少

2018 年 7 月 3 日提交;最初宣布 2018 年 7 月。

34. 第 1807. 08906[[pdf](#)] [cse](#)

[多伊](#) [1012s018539317400058](#)

基于关联规则挖掘的错误分配中冗余规则的减少

作者:[meera sharma](#), [abhishek tandon](#), [madhu kumari](#), [v b singh](#)

摘要: bug 试用是一个决定如何处理新出现的 bug 报告的过程。在本文中, 我们挖掘了使用不同的 bug 属性, 即严重性、优先级、组件和操作系统来预测新报告的 bug 受让人的关联规则。为了解决大数据集的问题, 我们采用 k-均值聚类算法, 通过划分大数据集, 对数据集进行了子集处理。我们在 matlab 中使用了 apriori 算法来生成关联规则。我们提取了每个集群中前 5 个受让人的关联规则。该方法已在 mozilla 开源软件项目(seamonkey、firefox 和 bugzilla) 的 14696 错误报告中得到了实证验证。该方法对现有的 bug 分配问题技术进行了改进。少

2018 年 7 月 23 日提交;最初宣布 2018 年 7 月。

评论:14 页

日记本参考:国际可靠性、质量和安全工程杂志,第 24 卷, 第 6 期 (2017) 1740005 (14 页)

35. 第 xiv: 1807. 08015[[pdf](#),其他] [cse](#)

uma análise 比较 de ferramentas de análise estática para detecção de erros de memória

作者:[patticia monteiro](#), [joao lourenço](#), [antónio ravares](#)

摘要: ---英文版软件故障往往与事故有关, 带来严重的经济和/或人类后果, 因此必须投资于软件的验证, 即关键的软件。本文通过对 c/c + 程序四种静态分析工具的可用性和有效性的比较分析, 探讨了软件质量的主题。这项研究使我们能够了解静态分析工具在软件验证和验证方面的巨大潜力和高度影响。作为一个补充结果, 在高受欢迎的开源程序中发现了新的错误, 并进行了报道。---英文版软件错误经常与事故有关, 造成严重的经济和/或人类后果, 因此必须对软件的验证进行投资, 即关键的软件验证。本文通过对 c/c + 程序的四个静态分析工具的可用性和效率进行比较分析, 探讨了软件质量的主题。这项研究可以了解这些工具在软件验证和验证方面可能产生的巨大潜力和高度影响。作为一个互补的结果, 我们在非常流行的开源项目中发现了新的错误, 这些错误已经被报道。少

2018 年 7 月 20 日提交;最初宣布 2018 年 7 月。

评论:葡萄牙文文章, 在全国信息学会议中被接受
(<http://inforum.org.pt/INForum2018>)

36. 第: 1807. 04728[[pdf](#),其他] [Cs](#)。直流

[多伊](#) [10.114/321940. 3219135](#)

科学令牌: 基于能力的远程科学数据安全访问

作者:[alex withers](#), [brian bockelman](#), [derek weitzel](#), [duncanbrown](#), [jeff gaynor](#), [jim basney](#), [todd tannenbaum](#), [zach miller](#)

摘要: 计算科学工作流的安全凭据 (例如密码、密钥) 的管理是科学家和信息安全官员的负担。凭据问题 (例如, 过期、特权不匹配) 导致工作流无法获取所需的输入数据或

存储有价值的科学结果, 从而通过要求科学家诊断问题、重新运行其计算, 并等待更长的时间, 等待他们的结果。在本文中, 我们介绍了科学令牌, 开源软件, 以帮助科学家更可靠和安全地管理他们的安全凭据。我们介绍了科学令牌系统的架构、设计和实现, 解决激光干涉仪引力波观测台 (ligo) 科学协作和大型天气测量望远镜 (lsst) 项目中的用例。我们还介绍了与支持分布式科学计算的广泛使用的软件的集成, 包括 htcondor、cvmfs 和 xrootd。本令牌使用 ietf 标准 oauth 令牌对远程科学数据进行基于功能的安全访问。访问令牌传达工作流所需的特定授权, 而不是通用身份验证模拟凭据, 以解决在分布式基础结构 (包括 nsf 资源) 上运行的科学工作流的风险 (例如, ligo 数据网格、开放科学网格、xsede) 和公共云 (例如, 亚马逊 web 服务、google 云、microsoft azure)。通过提高科学工作流的互操作性和安全性, 科学令牌 1) 支持将分布式计算用于需要更大数据保护的科学领域, 2) 通过减少更广泛的分布式计算资源, 实现了对更广泛的分布式计算资源的使用。远程系统上的凭据滥用的风险。少

2018 年 7 月 12 日提交;最初宣布 2018 年 7 月。

评论:8 页, 6 位数字, pearc ' 18: 高级研究计算的实践和经验, 2018 年 7 月 22-26 日, 匹兹堡, 宾夕法尼亚州, 美国

37. 第 xiv:1807. 04486[[pdf](#),[其他](#)] cse

特征选择对预测错误数量的影响

作者:[haidar osman](#), [mohammad ghafari](#), [oscar nierstras](#)

摘要: 错误预测是在软件度量和故障信息上训练机器学习模型以预测软件实体中的错误的过程。虽然特征选择是建立稳健预测模型的重要一步, 但没有足够的证据表明其对预测软件系统中错误数量的影响。我们研究了基于相关的特征选择 (cfs) 筛选方法和包装特征选择方法对五个广泛使用的预测模型的影响, 并演示了这些模型如何在选择不进行特征选择的情况下执行, 以预测五个不同的开源 java 软件系统中的错误。我们的结果表明, 包装的性能优于 cfs 过滤器;它们可将预测精度提高 33%, 同时消除一半以上的特征。我们还观察到, 尽管相同的要素选择方法在不同的项目中选择不同的要素子集, 但这个子集始终包含源代码和更改指标的组合。少

2018 年 7 月 12 日提交;最初宣布 2018 年 7 月。

38. 第 xiv:180. 7.04479[[pdf](#),[其他](#)] cse

rock: 使用众包知识在 ide 中进行代码搜索

作者:[mohammad masudur rahman](#), [chanchal k. roy](#), [david lo](#)

摘要: 传统的代码搜索引擎通常不能很好地使用自然语言查询, 因为它们大多应用关键字匹配。因此, 这些引擎需要精心设计的查询, 其中包含有关为代码搜索编程 api 编程的信息。遗憾的是, 现有的研究表明, 为代码搜索准备一个有效的查询对开发人员来说既具有挑战性, 也非常耗时。本文提出了一种新的代码搜索工具

--rack-rack-rack-rack-rack--- 它返回了用自然语言文本编写的给定代码搜索查询的相关源代码。该工具首先通过从堆栈溢出的众包知识中挖掘关键字 api 关联将查询转换为相关 api 类的列表, 然后将重新编写的查询应用于 github 代码搜索 api 以收集相关结果。提交与编程任务相关的查询后, 该工具会自动挖掘数千个开源项目中的相关代码段, 并将其显示为开发人员的编程环境-ide。工具页:

<http://www.usask.ca/~masud.rahman/rack> 少

2018 年 7 月 12 日提交;最初宣布 2018 年 7 月。

评论:第 39 届软件工程国际会议 (会议室卷) (icse 2017), 第 51-54 页, 阿根廷布宜诺斯艾利斯, 2017 年 5 月

报告编号: 10.1109/ICSE-C.2017.11

日记本参考:proc. icse-c 2017, 第 51-54 页

39. **建议: 1807. 03523**[pdf, ps,其他] Cs. Lg

深度学习优化库

作者:andrés camero, jamal toutouh, enrique alba

摘要: 深度学习超参数优化是一项艰巨的任务。寻找合适的网络配置是成功的关键, 然而, 这种劳动大多是大致完成的。在本文中, 我们介绍了一个新的库来解决这个问题, 深度学习优化库: dlpt。我们简要介绍了它的体系结构, 并给出了一组使用示例。这是一个在 gnu gpl v3 许可证下开发的开源项目, 可 <https://github.com/acamero/dlpt> 2018 年 7 月 10 日提交;最初宣布 2018 年 7 月。

评论:4 页

40. **第 xiv: 1807. 02965**[pdf,其他] cse

多伊 [10.11145/288916. 2889244](#)

正确: 基于跨项目和技术经验的 github 代码审阅者推荐

作者:mohammad masudur rahman, chanchal k. roy,jason a. collins

摘要: 对等代码评审在软件开发的早期阶段定位常见的违反编码规则和简单的逻辑错误, 从而降低总体成本。但是, 在 github 中, 为请求标识适当的代码审阅者是一项重要的任务, 因为评审器标识的可靠信息通常不容易获得。在本文中, 我们提出了一种代码评审器推荐技术, 该技术不仅考虑到相关的跨项目工作历史 (如外部库经验), 还考虑开发人员在某些专业技术方面的经验与一个拉请求, 以确定她作为一个潜在的代码审阅者的专业知识。我们首先使用对 10 个商业项目和 10 个相关库的探索性研究来激励我们的技术。使用来自 10 个商业项目和 6 个开源项目的 17 115 个请求的实验表明, 我们的技术提供了 85%-92% 的推荐精度, 约 86% 的精度和 79%-81% 的代码召回审查人员的建议, 这是非常有希望的。与最先进的技术相比, 也验证了我们的推荐技术的经验发现和优越性。少

2018 年 7 月 9 日提交;最初宣布 2018 年 7 月。

评论:第 38 届软件工程国际会议 (附录卷) (icse 2016), 第 222 页-231 页, 美国德克萨斯州奥斯汀, 2016 年 5 月

日记本参考:proc. icse-c 2016, 第 222 页--231

41. **第 xiv: 1807. 02261**[pdf,其他] cse

多伊 [10.1109/SCAM.2014.15](#)

试论上下文在推荐异常处理规范中的应用

作者:mohammad masudur rahman, chanchal k. roy

摘要: 研究表明, 软件开发人员通常要么滥用异常处理功能, 要么使用这些功能效率低下, 这种做法可能会导致正在进行的软件项目进入脆弱、不安全和不可靠的应用程序系统。在本文中, 我们提出了一种上下文感知代码推荐方法, 该方法建议从 github 托管的许多流行的开源代码存储库中获取异常处理代码示例。它收集利用 github 代码搜索 api 的代码示例, 然后通过不仅利用结构 (即基于图形) 和词汇功能, 还利用启发式质量, 对 ide 中正在开发的代码进行分析、筛选和排序示例中的异常处理程序的度

量值。对 4,400 个代码示例和 65 个异常处理方案进行的实验以及与现有四种方法的比较表明,该方法非常有希望。少

2018 年 7 月 6 日提交;最初宣布 2018 年 7 月。

评论:第 14 届 iee 源代码分析和操作国际工作会议 (scam 2014), 第 285 页-294 页, 加拿大维多利亚, 2014 年 9 月

日记本参考:proc. scam 2014, 第 285 页-294 页

42. 第 xiv:180.7.01769[pdf,其他] Cs. Ce

fluidsim: 模块化、面向对象的 python 包, 用于高性能 cfd 模拟

作者:ashwin vishnu mohan, cyrille bonamy, miguel calpe linares, pierre augier

文摘: 本文介绍了 python 包流体作为计算流体力学 (cfd) 求解器的可扩展框架。它是作为 FluidDyn 项目(augier 等人, 2018年) 的一部分开发的, 该项目旨在促进流体力学社区内的开源和开放科学合作, 既针对教育, 也适用于教育和科学合作。研究目的。

fluidsim 中的求解器是可扩展的高性能计算 (hpc) 代码, 由丰富、科学的 python 生态系统和 fluidsim 和 fluidfft 包提供的应用程序编程接口 (api) 提供的引擎盖下供电 (mohan 等人, 2018). 本文介绍了流体的设计方面, 即使用 python 作为主要语言;专注于代码的易用性、重用性和维护性, 而不会影响性能。简要介绍了实现的细节, 包括优化方法、功能的模块化组织和使用类实现求解器的面向对象方法。目前, 流体包括使用不同数值方法 (包括有限差分法) 解决各种物理问题的解算器。但是, 此元底纸应只涉及其伪光谱求解器的实现和性能, 特别是二维和三维纳维尔-斯托克斯求解器。我们研究了流体在最先进的 hpc 集群中的性能和可扩展性。提出了三种基于 python (dedalus, 光谱 dns) 和 fortran (ns3d) 的类似伪光谱 cfd 码, 并与流体进行了定性和定量。源代码作为一个汞存储库 bitbucket.org/fluidyn-fluidsim 托管在 bitbb 上, 使用 spinx 生成的文档可以在 fluidsim.readthedocs.io 在线阅读。少

2018 年 7 月 3 日提交;最初宣布 2018 年 7 月。

43. 第 xiv:807.00148[pdf,其他] cs. cy

房地产市场和交通的模块化微观模拟架构

作者:paul waddell, ignacio garcia-dorado, samuel m.maurer, geoff 波音公司, max gardner, emily porter, daniel aliaga

摘要: 整合土地使用、旅行需求和交通模式是区域规划的黄金标准, 但很少以有意义的方式实现, 特别是在分类数据的规模上。本文提出了一种新的城市土地利用、出行需求和交通分配的模块化微观仿真体系。urbansim 是一个开源的微仿真平台, 被世界各地的都市规划组织用于模拟城市在漫长 (约 30 年) 时间范围内的增长和发展。

activitysim 是一个基于代理的建模平台, 它基于 urbansim 模型和软件框架开发, 可生成合成的原始目的地旅行需求数据。对于流量分配, 我们集成了两种方法。第一种是静态用户均衡方法, 用作基准。第二种是流量微模拟方法, 我们在 gpu 上运行时具有高度并行性, 以便在整个建模工作流中实现代理的全模型微模拟。本文介绍了本课题的研究日程, 介绍了本项目迄今在开发该模块平台方面取得的成就, 并概述了进一步的研究方向。少

2018 年 6 月 29 日提交;最初宣布 2018 年 7 月。

评论:在应用城市建模研讨会上提交的论文。英国剑桥。2018 年 6 月 27-29 日。arxiv 管理说明: 与 arxiv:1802.09335 实质性文本重叠

44. 第 [xiv:1806.08457](#)[pdf,其他] cse

您要呼叫谁? : github 讨论中 @-菜单的决定因素

作者:[david kavaler](#), [premkumar devanbu](#), [vladimir filkov](#)

摘要: 开源资料来源软件 (oss)项目的成功取决于人群的贡献。当基于拉请求的系统中出现问题时, @ 提及被用来号召人们进行任务;以前的研究表明, 讨论中的 @ 提及与更快的问题解决方法有关。在大多数项目中, 可能有许多开发人员在技术上能够处理各种任务。但 oss 支持分布在广泛的社会和地理背景以及参与程度的动态团队。那么, 重要的是要知道该找谁, 即谁可以被依赖或信任承担与任务有关的重要职责, 以及原因。在本文中, 我们试图了解开发人员的哪些可观察到的社会技术属性可以用来构建一个很好的模型, 即它们在 github 问题中被提及, 并拉请求讨论。我们构建了未来 @ 提及的整体和项目特定的预测模型, 以便在 200 个 github 项目中的每一个项目中捕获 @ 提及的决定因素, 并了解这些决定因素之间是否以及如何不同项目。我们发现, 可见性、专业知识和生产率与 @ 提及的增加有关, 而在许多控制变量的存在下, 响应能力并不高。此外, 我们发现, 虽然存在项目特定的差异, 但整个模型可用于跨项目预测, 这表明了其 github 范围的实用程序。少

2018 年 6 月 21 日提交;最初宣布 2018 年 6 月。

评论:12 页, 5 个数字, 2 个表

类:D.2。2

45. 第 [1806.08149](#)[pdf,其他] cse

开发人员对过时和违反许可证的堆栈溢出代码的认识和经验: 在线调查

作者:[chaiyong Raghitwetsagul](#), [jens kim k 皇后](#), [rocco ololitolet](#)

摘要: 我们对堆栈溢出应答器和访问者进行了两次在线调查, 以评估他们对堆栈溢出应答器中过时的代码和软件许可证的认识。答复者调查针对 607 名享有盛誉的堆栈溢出用户, 并获得 33% 的高回复率。我们的发现如下。尽管答案中的大多数代码段都是从头开始编写的, 但也有一些代码段是从相应的问题、个人或公司项目或开源项目克隆出来的。堆栈溢出应答者知道他们的一些片段已经过时。然而, 19% 的参与者报告说, 他们很少或从来没有修复他们过时的代码。至少 98% 的应答者从不在其代码段中包含软件许可证, 69% 的答复者从不检查与 stack 溢出的 cc bi-sa 3.0 的许可冲突, 如果他们将代码从其他来源复制到堆栈溢出答案。游客调查使用了方便的抽样, 收到了 89 份答复。我们发现 66% 的参与者经历了克隆和重用堆栈溢出片段的问题。56% 的游客从未向堆栈溢出帖报告过这些问题。85% 的参与者不知道 stack 溢出应用了 cc by-sa 3.0 许可证, 62% 的参与者从不将代码复制给堆栈溢出的帖子或答案归因。此外, 66% 的参与者不检查复制的堆栈溢出代码与其软件之间的许可冲突。通过这些发现, 我们建议堆栈溢出提高他们的用户 (包括回答者和访问者) 对过时和违反许可证的代码段问题的认识。少

2018 年 6 月 21 日提交;最初宣布 2018 年 6 月。

评论:存档的附加材料: 1806. 07659

报告编号:rn 17/

46. 第 [xiv:1806.08126](#)[pdf,其他] Cs。Dm

利用拓扑工具包简化拓扑数据分析

作者:guillaume favelier, charles gueunet, attila gyulassy, Gyulassy
kitware, joshua levine, jonas lukasczyk, daisuke sakurai, maxime soler, Gyulassy
tiemy, 威尔·厄谢尔, 齐武

摘要: 本教程介绍了从用户的角度分析和可视化科学数据的拓扑方法, 使用拓扑工具包 (ttk), 这是一个最近发布的用于拓扑数据分析的开源库。在过去的二十年里, 拓扑方法的普及和成熟程度有了相当大的提高, 既定方法的成功案例也被记录在广泛的应用中 (燃烧、化学、天体物理学、材料科学等)。在事后和现场的情况下获取和模拟数据。虽然已经出版了关于这个主题的参考教科书, 但 *ieee vis* 的任何教程近年来都没有涉及这一领域, 从来没有从软件层面和用户的角度来看。本教程通过为从业者、研究人员、学生和讲师提供对拓扑方法的初学者介绍来填补这一空白。特别是, 本教程没有关注理论方面和算法细节, 而是重点介绍拓扑方法如何在实际中对具体的数据分析任务 (如分割、特征提取或跟踪) 有用。本教程详细介绍了如何使用 *ttk* 完成这些任务。首先, 在介绍了拓扑方法及其在数据分析中的应用之后, 将简要概述 *ttk* 为最终用户提供的主要切入点, 即 *pareview*。其次, 将概述 *ttk* 的主要特点。将详细介绍一个正在运行的示例, 演示如何通过 *pareview*、*python*、*vtksc++* 和 *c++* 访问 *ttk* 的功能。第三, 实践课程将具体展示如何在 *pareview* 中使用 *ttk* 来执行多个具有代表性的数据分析任务。第四, *ttk* 的使用将为开发人员提供, 特别是通过描述在 *ttk* 之上构建的可视化和数据分析项目的几个示例。最后, 对 *ttk* 作为拓扑分析教学平台的使用提出了一些反馈意见。本教程的主讲人包括拓扑方法方面的专家、*ttk* 的核心作者以及来自学术界、实验室或行业的活跃用户。本教程的很大一部分将专门用于动手练习, 并将向参与者提供丰富的材料包 (包括在虚拟机中预装 *ttk*、代码、数据、演示、视频教程等)。本教程主要针对的是不是拓扑方法专家但有兴趣在日常工作中使用这些方法的学生、从业者和研究人员。我们还针对已经熟悉拓扑方法的研究人员, 他们有兴趣使用或为 *ttk* 做出贡献。少

2018 年 6 月 21 日提交;最初宣布 2018 年 6 月。

47. 第 [xiv:866.07659](#)[pdf,其他] cse

堆栈溢出上的毒性代码段

作者:chaiyong Ragkhitwetsagul, jenskoinke, matheus paixao, giuseppe bianco, rocco ololaol 论

摘要: 联机代码克隆是从软件项目或联机源复制到堆栈溢出作为示例的代码片段。由于在代码复制到堆栈溢出后没有检查机制, 它们可能会成为有毒的代码段, 即它们会导致过时或违反原始软件许可证。我们通过结合两个开发人员调查和大规模代码克隆检测, 对堆栈溢出及其毒性的在线代码克隆及其毒性进行了研究。对 201 名高信誉堆栈溢出应答器 (33% 的回复率) 进行的调查显示, 131 名参与者 (65%) 曾收到过时代码的通知, 其中 26 人 (20%) 很少或从未修复过代码。138 答复者 (69%) 从未检查过许可证之间的冲突。他们复制的代码段和堆栈溢出的 *cc by-sa 3.0*。对 87 堆栈溢出访问者的调查显示, 他们遇到了来自堆栈溢出答案的几个问题: 不匹配的解决方案、过时的解决方案、不正确的解决方案和错误代码。85% 的用户不知道堆栈溢出强制使用 *cc bi-sa 3.0* 许可证, 66% 的用户在重用代码段时从不检查许可证冲突。我们的克隆检测发现在线克隆对在堆栈溢出上的 72,365 java 代码段和 111 个开源项目在管理的 *qualitas* 语料库中。我们分析了 2,289 名非平凡的在线克隆候选。我们的调查显示, 有有力的证据表明, 153 个克隆被从 *qualitas* 项目复制到堆栈溢出。我们发现其中 100 项

(66%) 已经过时, 可能对再利用有害。此外, 我们发现 214 个代码段可能违反其原始软件的许可, 并在 2, 427 次 github 项目 中出现 7, 112 次。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

48. 第 [xiv:866.07619](#)[pdf,其他] cse

多伊 [10.1109/SEAA.2018.00068](#)

源代码审阅者推荐的大规模研究

作者:[jakub lipcak](#), [bruno rossi](#)

摘要: 上下文: 软件代码评审是开发过程的重要组成部分, 可提高软件质量并降低总体成本。但是, 找到合适的代码审阅者是一项复杂且耗时的任务。目标: 在本文中, 我们提出了一个大规模的研究, 以比较两个主要的源代码审阅者推荐算法 (revfinder 和一个朴素的贝氏方法) 在识别打开的拉的最佳代码审阅者的性能请求。方法: 我们从 github 和 Gerrit 存储库中挖掘数据, 构建一个由 51 个项目组成的大型数据集, 分析了超过 293k 的拉请求, 180k 所有者和 157k 评审员。结果: 根据大量的分析, 我们可以说, (一) 任何模型都不能推广为最好的所有项目, (二) 使用不同的存储库 (Gerrit, github) 可以对推荐结果产生影响, (三) 利用子项目在 Gerrit 中提供的信息可以改进推荐结果。少

2018 年 6 月 20 日提交;最初宣布 2018 年 6 月。

评论:在第 44 届欧洲软件工程和高级应用会议 (seaa 2018) 上发表

49. 第 [1806.06143](#)[pdf,其他] Cs. 佛罗里达州

多伊 [10.4230/LIPIcs.CONCUR.2018.20](#)

选择性监控

作者:[radu grigore](#), [Kiefer kiefer](#)

摘要: 我们研究了标记马尔可夫链的选择性监视器。监视器观察马尔可夫链在运行过程中产生的输出, 目的是将运行标识为正确或错误。如果监视器跳过观察以减少监视开销, 则它是选择性的。我们对最大限度地减少预期观测次数的监视器感兴趣。我们建立了一个不可判定的结果, 用于有选择地监测一般马尔可夫链。另一方面, 对于非隐藏马尔可夫链 (任何输出标识马尔可夫链的状态), 我们展示了简单的最优监视器的存在, 并且可以基于 dfa 语言等价性进行有效计算。这些监视器不依赖于马尔可夫链中的精确转换概率。我们报告了一些开源 java 项目的计算这些监视器的实验。少

2018 年 7 月 2 日提交;v1 于 2018 年 6 月 15 日提交;最初宣布 2018 年 6 月。

评论:2018 年

50. 第 [xiv:1806.04556](#)[pdf, ps,其他] cse

基于开放式项目的大规模研究评价维修性偏见

作者:[tobias roehm](#), [daniel Veihelmann](#), [stefan wagner](#), [elmar juergens](#)

摘要: 夸张或上下文更改可能会使可维护性体验陷入偏见。例如, javascript 通常被视为最不优雅的语言, 因此具有最低的可维护性。这种偏见不应在没有事先经验验证的情况下指导决定。我们根据偏见制定了 10 个关于可维护性的假设, 并在大量的开源项目 (6, 897 github 存储库、4.0 亿行、5 种编程语言) 中对其进行了测试。我们使用五个静态分析指标来实现可维护性。我们发现 javascript 代码并不比其他代码差, java 代码比 c# 代码具有更高的可维护性, c 代码的方法比其他代码更长。在 java 代码中, 接口文档的质量比在其他代码中更好。团队开发的代码不是较高和较大的代码库, 也不是

较低的可维护性。具有高可维护性的项目并不更受欢迎,也不经常分叉。总体而言,大多数假设都没有得到开源数据的支持。少

2018 年 6 月 12 日提交;最初宣布 2018 年 6 月。

评论:20 页

51. 第 [xiv:1806.02592](#)[pdf,其他] cse

一种基于 nlp 的支持开源社区入职和保留的简单方法

作者:[christoph stanik](#), [lloyd montgomery](#), [Stanik martens](#), [davidde fucci](#), [walid maalej](#)

摘要: 成功的开源社区不断寻找新成员,并帮助他们成为活跃的开发人员。开发人员在开源项目中加入的一种常见方法是让新来者专注于相关但易于解决的问题,以熟悉代码和社区。这项研究的目标有两个方面。首先,我们的目标是通过分析已解决问题的历史,简单地使用问题的标题和描述,自动识别新来者可以解决的问题。其次,我们的目标是自动识别问题,这些问题可以由后来成为活跃开发人员的新来者来解决。我们挖掘了三个大型开源项目的问题跟踪器,并从已解决问题的标题和描述中提取了自然语言特征。在一系列实验中,我们对四个监督分类器的准确性进行了优化和比较,以满足我们的研究目标。在第一个进球时,随机森林的精度达到了 91% (f1-分数 72%),而对于第二个进球,诊断树的精度达到了 92% (f1-分数 91%)。定性评估使人们深入了解问题描述中的哪些信息对新来者有帮助。我们的方法可用于仅基于问题的文本自动识别、标记开源软件项目中的新来者,并为他们推荐问题。少

2018 年 8 月 16 日提交;v1 于 2018 年 6 月 7 日提交;最初宣布 2018 年 6 月。

52. 第 [xiv:1906.00802](#)[pdf,其他] 反渗透委员会

maestro: 用于低层控制和高级推理的集成编排的机器人框架

作者:[asim munawar](#), [giovanni de r](#) 秋护酒店, [Tu-Hoapham](#), [daiki kim ura](#), [michiaki tatsubori](#), [takao moriyama](#), [ryuki tachibana](#), [grady booch](#)

摘要: 本文介绍了一个叫做 maestro 的框架。它的设计是通过自然语言或演示给出的简单的高级指令,使机器人能够高精度地完成复杂的任务。为了实现这一点,它通过使用存储在本体形式和规则中的知识来处理层次结构,以便在不同级别的指令之间进行桥接。因此,该框架有多层处理组件;低级别的感知和驱动控制,符号规划器和 watson api 的认知能力和语义理解,以及这些组件的编排由一个新的开源机器人中间件称为 project 因图的核心。我们展示了如何在多个参与者(人、通信机器人和工业机器人)协作执行共同工业任务的复杂场景中使用此框架。人类用自然语言对话和演示向 pepper (软银机器人的人形机器人) 教授组装任务。我们的框架帮助 pepper 感知人类演示,并为 ur5 (来自通用机器人的协作机器人手臂) 生成一系列动作,最终执行组装(例如插入)任务。少

2018 年 6 月 3 日提交;最初宣布 2018 年 6 月。

评论:ieee 机器人与自动化国际会议 (icra) 2018。视频:

<https://www.youtube.com/watch?v=19JsdZi0TWU>

53. 第 [1805.5.12398](#)[pdf] cs. cy

多伊 [10.1007/s10606-018-9333-1](#)

数据分析开源软件库中文档的类型、作用和实践:文档工作的协作民族志

作者:[r. stuart geiger](#), [nelle varoquaux](#), [charlotte mazel-cabasse](#), [chris holdgraf](#)

摘要: 计算研究和数据分析越来越依赖于开源 软件 (oss) "库" 的复杂生态系统----这些库是程序员为执行特定任务而导入的可重用代码的集合。这些库的软件文档对于帮助程序员/分析师了解哪些库可用以及如何使用这些库至关重要。然而,开源 软件库的文档被广泛认为是低质量的。本文是 cscw 研究人员和贡献者之间的合作数据分析 oss 图书馆, 基于人种学实地工作和定性访谈。我们研究了围绕这些主要以志愿者为基础的项目中的格式、实践和文档方面的挑战的几个问题。这些库周围存在着许多不同类型和格式的文档, 它们发挥着各种教育、宣传和组织作用。文档背后的工作同样是多方面的, 包括编写、审查、维护和组织文档。文件工作的不同方面要求贡献者拥有不同的技能, 克服各种社会和技术障碍。最后, 我们的大多数受访者并不报告做文档工作的内在享受程度很高 (与编写代码相比)。他们的动机受到个人和项目特定因素的影响, 例如, 与添加新功能或修复 bug 等更多 "技术" 任务相比, 执行文档工作的信用水平达到了一定程度。在研究数据分析 oss 库的文档工作时, 我们获得了一个新的窗口, 了解数据密集型研究不断变化的实践, 并帮助从业人员更好地了解如何支持这些通常是无形的基础设施工作。项目。少

2018 年 5 月 31 日提交;最初宣布 2018 年 5 月。

日记本参考:计算机支持的合作工作。2018. doi:10.100/10606-018-9333-1

54. 第 1805 5.1651 [pdf,其他] Cs. CI

利用双向 lstm 递归神经网络拆分源代码标识符

作者:vadim markovtsev, waren long, egor bulychyev, romain keramitas, konstantin slavnov, gbor markowski

摘要: 程序员在他们通过标识符和注释编写的源代码中充分使用自然语言。源代码标识符是从与含义、命名约定和上下文密切相关的令牌池中选择。这些令牌通常结合在一起, 以产生更精确和更明显的指定。此类多部分标识符占公共 git 存档中所有命名令牌的 97%, 这是迄今为止 git 存储库的最大数据集。我们引入了双向 lstm 递归神经网络来检测源代码标识符中的子令牌。我们对该网络进行了 41.7 种不同的拆分标识符的培训, 这些标识符是从 public git 存档中的 1, 014 个开源项目中收集的, 并表明它优于其他几种机器学习模型。建议的网络可用于改进基于源代码标识符的上游模型, 以及改进开发人员在不切换键盘大小写的情况下编写代码的体验。少

2018 年 7 月 19 日提交;v1 于 2018 年 5 月 26 日提交;最初宣布 2018 年 5 月。

评论:8 页

55. 第 xiv:1805 5.723 [pdf,其他] Cs. Hc

民主化设计: 通过制造商套件介绍人工智能的新情况

作者:victor dibia, maryam ashoori, aaron cox, justin weisz

摘要: 现有的研究突出了当技术充分民主化并使非技术或新手用户能够获得时所实现的无数好处。然而, 人工智能 (ai) 等复杂技术的民主化仍然很困难。在这项工作中, 我们借鉴了创新民主化的理论基础, 探索制造商套件的设计, 帮助新手用户向复杂的技术。我们报告了我们设计 tjbot 的工作: 一种开源 纸板机器人, 可以使用预先构建的 ai 服务进行编程。我们重点介绍了我们在此过程中采用的原则 (平易近人的设计、简单性、可扩展性和可访问性)、我们从研讨会上展示套件 (66 人参加) 中学到的见解, 以及用户如何通过 12 个月期间 (2016 年 11 月-2017 年 11 月)。我们发现, 该项目成功地吸引了新手用户 (40% 的分叉项目的用户是 github 的新手), 各种人口统计对原型化用例感兴趣, 如家庭自动化、任务委派、教学和学习。少

2018 年 7 月 5 日提交;v1 于 2018 年 5 月 27 日提交;最初宣布 2018 年 5 月。

评论:早期的文件草稿-更新的参考资料, 数字说明

56. 第 1805 5.10511[[pdf](#), [ps](#),其他] Cs. Db

以实用型模式挖掘为研究综述

作者:陈文生,林春伟, 菲利普 fownier-viger, han-chieh chao, vincent s.tseng, phillip s. yu

摘要: 数据挖掘和分析的主要目的是找到新的、潜在有用的模式, 这些模式可在现实世界的应用程序中使用, 以获得有益的知识。为了识别和评价不同类型模式的效用, 提出了许多技术约束, 如支持、置信度、顺序排序和效用参数 (如重量、价格、利润、数量等)。近年来, 对公用事业型模式挖掘 (upm) 的需求越来越大。upm 是一项至关重要的任务, 拥有众多影响大的应用, 包括跨营销、电子商务、金融、医疗和生物医学应用。本调查旨在全面、全面和结构化地概述 upm 的最新方法。首先, 我们介绍了对 upm 的深入了解, 包括概念、示例以及与相关概念的比较。介绍了挖掘不同类型的高实用新型的最常见和最先进的方法的分类, 包括基于 Apriori-based、基于树的、基于投影的、基于垂直/水平数据格式的方法和其他混合方法。方法。本文对现有的高实用模式挖掘技术的高级主题进行了全面回顾, 并对其利弊进行了讨论。最后, 我们为 upm 提供了几个著名的开源软件包。在调查结束时, 我们讨论了这一领域的公开和实际挑战。少

2018 年 5 月 26 日提交;最初宣布 2018 年 5 月。

57. 决议: 1805 5.09251[[pdf](#),其他] Cs. 镍

apmec: 用于多址边缘计算的自动资源调配框架

作者:tong v. doan , giang t. nguyen, 亚历山大 kropp , frank h. p. fitdek

摘要: 5g 和触觉互联网领域的新型使用案例和垂直领域, 如互联汽车和人机合作, 可从网络功能虚拟化 (nfv) 和多访问边缘提供的灵活性和减少延迟中大大受益计算 (mec)。管理和协调 mec 和 nfv 的现有框架要么紧密耦合, 要么完全分离。前一种设计是不灵活的, 增加了一个框架的复杂性。而后者导致计算资源的使用效率低下, 因为信息没有共享。我们引入了 apmec, 这是一个专门的 mec 框架, 同时支持与 nfv 的管理和编排 (mano) 框架的协作。新的设计允许重用分配的网络服务, 从而最大限度地提高资源利用率。测量结果表明, apmec 可以分配多达 60% 以上的网络服务。apmec 是在 openstack 之上开发的, 是一个开源项目, 可用于协作和促进进一步的研究活动。少

2018 年 5 月 9 日提交;最初宣布 2018 年 5 月。

58. xiv:1805.06427[[pdf](#),其他] Cs. Hc

多伊 [10.1088/1741-2552/aada0](#)

mobbb: 基于可信的 bci 算法基准

作者:vinay jadaram , 亚历山大·巴拉昌特

摘要: bci 算法的开发长期以来一直受到两个主要问题的阻碍: 小样本集和缺乏重现性。我们通过软件套件为这两个问题提供解决方案, 该解决方案简化了以可靠方式查找和预处理数据的问题, 以及使用一致的机器学习方法接口的问题。在 mne 工具包中实现的信号分析软件的最新进展和科学学习项目提供的机器学习统一框架的基础上, 我们提供了一个可以改进 bci 算法开发的系统。该系统是 bsd 许可证下的完全开源的, 可在 <https://github.com/NeuroTechX/moabb> 上使用。为了验证我们的努力, 我们分析

了一组跨 12 个开放访问数据集的最先进的解码算法, 其中 250 多个主题。我们的分析证实, 不同的数据集可能会导致相同的处理管道非常不同的结果, 这突出表明需要在 bci 领域建立可信的算法基准, 而且许多以前验证过的方法不会跨不同的数据集进行应用时, 这对实际的 bci 具有广泛的影响。少

2018 年 5 月 16 日提交;最初宣布 2018 年 5 月。

59. 第 1805.06297[[pdf](#),其他] Cs. CI

一种用于完全无监督的单词嵌入跨语言映射的可靠自学习方法

作者:mikel artetxe, gorka labaka, eneko agirre

摘要:最近的工作通过对抗训练将单语嵌入映射到共享空间, 成功地学习了没有平行数据的跨语言单词嵌入。然而, 他们的评估侧重于有利条件, 使用可比的语料库或密切相关的语言, 我们表明, 他们往往失败在更现实的情况下。本工作提出了一种基于完全无监督初始化的替代方法, 该方法明确利用了嵌入的结构相似性, 并提出了一种鲁棒的自学习算法, 迭代地改进了该解决方案。我们的方法在所有测试方案中都取得成功, 并在标准数据集中获得最佳发布结果, 甚至超过以前的监督系统。我们的实施是作为一个开源项目发布的 <https://github.com/artetxem/vecmap> 少

2018 年 5 月 17 日提交;v1 于 2018 年 5 月 16 日提交;最初宣布 2018 年 5 月。

评论:acl 2018

60. 第 1805.0 04457[[pdf](#)] cs. cy

基于游戏开发的学习体验: 游戏中的性别差异

作者:bernadette spieler, wolfgang slani

摘要:学习理论强调内在和外在动机在课程中的重要性, 游戏是一种很有希望的方式, 在构建游戏和在公开场合或与社区展示或分享游戏时既提供又一个。新技术和新兴的手机游戏行业进一步证明, 应该随时随地促进学习。这似乎是所有学生以娱乐方式学习编码的一个有希望的机会, 这就提出了一个问题, 即这种基于游戏的概念是否也有助于解决妇女在信息技术相关领域的性别差距。中学中已经存在性别差异。这些年是第一次职业选择, 但对技术科目的参与程度也很低的年份。为了解决这种性别偏见, 欧洲项目"不让一个人掉队" (nolb) 的目标是整合袖珍代码, 这是由 catrobat 开发的一个应用程序, 这是技术大学的一个免费开源 非营利项目。格拉兹/奥地利, 进入不同的学校科目, 从而使编码更容易接触和吸引女生。在本项目期间 (2015-2017 年), 支持教师通过构建想法和通过游戏设计实现理念, 指导学生学习过程。本文根据提交的程序的游戏设计对其提交的程序进行了分析。在细节上, 评估考虑了正式的游戏元素, 游戏结构, 和使用的图形, 以及袖珍代码的具体方面。程序分析显示了性别常用的设计模式, 并提出了女性青少年首选的游戏设计特征。这一分析有助于建立一个更具创造性和包容性的基于游戏开发的学习 (gdbl) 环境, 通过构建个人体验, 在内在和外在激励的基础上, 为自我表达和激励提供空间。少

2018 年 8 月 31 日提交;v1 于 2018 年 5 月 11 日提交;最初宣布 2018 年 5 月。

评论:10 页, 8 位数字, ecgbl 2018

日记本参考:2018 年欧洲游戏学习大会

61. 第 1805.0 4366[[pdf](#)] cs. cy

女性青少年与编码: 创造对性别敏感和创造性的学习环境

作者:bernadette spieler, wolfgang slani

摘要: 妇女在技术领域的人数远远低于男性的平均人数,特别是在发达国家。在 12 至 15 岁的学生中,中学中已经存在 stem 中的性别差异。青春期是身份形成的关键时刻,自我属性是内部冲突的**根源**,特别是对女性青少年而言。正是在这一中间女性青春期,女孩开始作出关键的职业选择,因此,这是加强她们并缩小信息和通信技术方面性别差距的关键时代。从哲学的角度来看,计算思维能力很重要,因为它们使我们能够以清晰、容易理解、但又鼓舞人心和具有挑战性的方式理解理性思维的基础。为了解决学校中的性别偏见问题,欧洲 h2020 项目的目标之一是整合口袋代码,这是一个由非营利项目 catrobat 开发的**免费开源** 应用程序,到不同的学校课程。通过游戏设计,袖珍代码允许少女融入多样性和包容性,以及反映其文化身份、情绪、喜好以及互动和思维方式的能力。为了评估在这些课程中使用应用程序的影响,我们捕获了让女孩参与设计和编码活动的结果。在本文中,作者介绍了**项目第二周期**的调查数据。以女性青少年为重点,研究结果让我们得出结论,合适的课堂环境对她们来说比编码工具本身重要得多。

少

2018 年 8 月 31 日提交,v1 于 2018 年 5 月 11 日提交;最初宣布 2018 年 5 月。

评论:12 页, 8 位数字, 2018 年建构主义

日记本参考:建构主义 2018

62. 第 1805.503677[[pdf](#)] Cs. Db

数据集营养标签: 推动更高数据质量标准的框架

作者:[sarah holland](#), [ahmed hosny](#), [sarah newman](#), [joshua joseph](#), [kasia chmielinski](#)

摘要: 建立在不完整或有偏见的数据基础上的人工智能 (ai) 系统往往会显示出有问题的结果。目前的数据分析方法,特别是在模型开发之前,成本很高,没有标准化。数据集营养标签 (标签) 是一个诊断框架,通过在 ai 模型开发之前提供对 "数据集成分" 的提炼而全面的概述,降低了标准化数据分析的障碍。构建可跨域和数据类型应用的标签要求框架本身具有灵活性和适应性;因此,标签由通过多个统计和概率建模后端生成的各种定性和定量模块组成,但以标准化格式显示。为了演示和推进这一概念,我们生成并发布了一个**开源** 原型,其中包含文档数据集的 propublica 美元上的七个示例模块。标签的好处是多方面的。对于数据专家来说, label 将推动更可靠的数据分析实践,提供一种有效的方法来为其目的选择最佳数据集,并通过更强大的培训数据集和检查能力来提高 ai 模型的整体质量模型开发时的问题。对于那些构建和发布数据集的用户, "标签" 创建了对解释的期望,这将推动更好的数据收集实践。我们还探讨了标签的局限性,包括跨不同数据集进行泛化的挑战,以及使用 "地面真相" 数据作为比较数据集的风险。鉴于所确定的限制,我们讨论了前进的道路。最后,我们为数据集营养标签项目指明了未来的方向,包括研究和公共政策议程,以进一步推动对这一概念的审议。

少

2018 年 5 月 9 日提交;最初宣布 2018 年 5 月。

评论:2018 年 5 月初稿

63. 第 1805.0051965[[pdf](#),[其他](#)] cse

描述 1000 个以上 java 批注的用法和影响

作者:[余忠兴](#),[白成刚](#), [莱昂内尔·塞恩图瑞尔](#),[马丁·蒙佩勒斯](#)

摘要: 注释作为一种元数据功能,自 java 5 以来已正式引入 java。自我介绍以来,java 社区已将注释广泛用于不同的目的,例如编译器指南和运行时处理。尽管使用量越来越多,但关于开发人员在实践中如何使用注释以及注释使用对软件质量的影响的经验证

据仍然有限。为了填补这一空白,我们对 github 上的 1,094 个开源项目进行了首次关于 java 注释使用的大规模实证研究。我们的研究回答了一些有关 java 注释使用的基本问题。首先,我们回答 "注释什么?" 的问题,并发现注释通常用于注释程序元素的 4 个方面。其次,我们回答 "注释是如何演变的" 的问题,并确定了 6 种不同的注释变化类型、它们的频率和特征。第三,我们回答 "谁使用批注" 的问题,并建立批注使用与代码所有权和开发人员体验之间的关系。此外,我们还使用回归模型来探讨注释使用与代码质量之间的相关性,我们发现注释确实会对减少代码出错的影响。少

2018 年 5 月 4 日提交;最初宣布 2018 年 5 月。

评论:25 页, 9 个数字, 10 个表

64. 第 1805.501342[[pdf](#),其他] cse

世界开源发展的比较研究

作者:泰国人 [mombach](#), [marco tulio valente](#), [cuiting chen](#), [magier bruntink](#), [gustavo pinto](#)

摘要: 开源软件在我们的现代社会中越来越重要, 为其他软件系统提供基本服务, 并支持各种最终用户应用程序的快速发展。最近, 全球代码共享平台 (如 github) 也在为开源的增长做出贡献。然而, 人们对这种增长在世界各地的分布情况以及不同国家制定的项目的特点了解甚少。在本文中, 我们提供了在 20 个国家开发的 2,648 个开源项目的描述。我们揭示每个国家的项目数量、每个国家项目的受欢迎程度和编程语言, 并显示一个国家的项目数量如何与其国内生产总值相关。最后, 我们使用一个名为 bettercodehub 的工具来评估所研究项目的可维护性和内部代码质量。少

2018 年 5 月 3 日提交;最初宣布 2018 年 5 月。

评论:11 页, 8 页

65. 第 1805.0132[[pdf](#),其他] cse

多伊 [10.114/333440.3195022](#)

海报: 低故障风险方法的识别

作者:[rainer nierdemayr](#), [tobias röhms](#), [stefan wagner](#)

摘要: 测试资源通常是有限的, 因此通常不可能在发布之前完全测试应用程序。因此, 测试人员需要将其活动重点放在相关的代码区域上。本文介绍了一种反缺陷预测方法来识别几乎不含任何故障的方法。我们将我们的方法应用于六个 java 开源项目, 并表明平均 31.6% 的项目方法具有较低的故障风险;它们总共只包含所有故障的 5.8%。此外, 结果表明, 与缺陷预测不同, 我们的方法也可以应用于跨项目预测场景。因此, 反向缺陷预测可以帮助确定未测试代码区域的优先级, 并指导测试人员提高故障检测概率。少

2018 年 5 月 3 日提交;最初宣布 2018 年 5 月。

评论:icse 2018 海报跟踪

类:D.2.5;D.2。 8

期刊参考: 2018 ieee ic 软件工程伙伴国际会议 (icse 伴侣)

66. 第 xiv:1804.10788[[pdf](#)] Cs. 镍

基于传感器的半持久调度在 c-v2x 网络中的性能分析

作者:[amr nabil](#), [vuk marojevic](#), [komalbir kaur](#), [carl dietrich](#)

摘要: 第三代合作伙伴项目在第 14 版中发布了蜂窝到所有 (c-v2x) 规范, 作为 lte 框架的一部分。c-v2x 是专用短距离通信的替代方案, 两者都是专为 v2x 控制信令而设计的。c-v2x 通过增加另外两种针对车辆环境的操作模式来扩展设备到设备的规格, 以覆盖 lte 基站为覆盖范围。车辆对车辆通信 (v2v) 是在模式 4 中建立的, 在模式 4 中, 设备采用基于传感器的半持久性调度 (sps), 以分布式方式安排传输。需要进行研究来评估 sps 的性能, 特别是在拥挤的无线电环境中。本文介绍了第一个能够进行这种研究的开源 c-v2x 模拟器。模拟器在 ns-3 中实现。分析了模式 4 资源池配置和一些关键 sps 参数对调度性能的影响, 发现资源预留间隔对数据包数据速率性能有显著影响, 而资源重新选择则对数据包数据速率性能有显著影响概率在密集的车辆高速公路场景中影响不大。结果表明, 正确配置调度参数可以显著提高性能。我们的结论是, 需要对拥塞控制机制进行研究, 以进一步提高许多实际使用案例的 sps 性能。少

2018 年 4 月 28 日提交;最初宣布 2018 年 4 月。

67. 第 xiv:1804.04621[[pdf](#),其他] cse

java 构建框架: 大规模编译

作者:[pedro martins](#), [rohan achar](#), [cristina v. lopes](#)

摘要: 大型的源代码存储库的研究往往将其效用限制在代码的静态分析上, 因为它们不能保证项目是否可以编译,更不用说以任何方式运行了。缺乏大型可编译和可运行数据集的直接后果是, 需要此类属性的研究不会超出小型基准的推广范围。我们提出了 java build 框架, 这是一种能够自动编译大量 java 项目的方法和工具, 这些项目可在诸如 github 这样的开源存储库中使用。核心是两个元素: 一个非常大的 jar 文件存储库, 以及解决编译错误和依赖关系的技术。少

2018 年 4 月 12 日提交;最初宣布 2018 年 4 月。

68. 第 1804.03589[[pdf](#),其他] cse

多伊 [10.1109/TSE.2018.2791521](#)

预测器: 现实世界中的并发缺陷预测

作者:[余廷婷](#),[魏文](#),[薛汉](#),[简·海斯](#)

摘要: 并发程序由于其固有的非确定性而难以测试。为了解决这个问题, 测试通常需要探索程序的线程计划;这在应用于实际程序时可能非常耗时。软件缺陷预测已被用来帮助开发人员查找错误并确定其测试工作的优先级。以前的研究使用机器学习来建立这样的预测模型, 基于设计的功能, 编码程序的特点。然而, 研究的重点是顺序方案;到目前为止, 还没有工作考虑过并发程序的缺陷预测, 程序特征与顺序程序不同。本文介绍了 con 副 tor, 一种通过静态和动态程序度量相结合来预测并发程序特有缺陷的方法。具体来说, 我们提出了一套基于并发程序的独特属性的新的静态代码度量。我们还利用基于突变分析构建的动态指标的其他指导。我们对四个大型开源项目的评价表明, 与传统特征相比, con 预测器改进了项目内缺陷预测和跨项目缺陷预测。少

2018 年 4 月 10 日提交;最初宣布 2018 年 4 月。

评论:这项工作得到了国家安全基金赠款 ccf-1464032、ccf-1652149 和 ccf-1511117 的部分支持

69. 第 1804.03518[[pdf](#),其他] cs. cy

多伊 [10.114/39477](#) [9.3194787](#)

千禧一代的软件工程

作者:jocelyn simmonds, maíra marques samary, milenko 竞技 omic, francisco madrid, con 斯坦扎·escobar

摘要: 软件工程师需要管理技术和专业技能才能取得成功。我们的大学提供了一个 5.5 年的课程, 混合了计算机科学, 软件和计算机工程, 其中前两年主要是数学和物理课程。因此, 我们的学生的第一个真正的团队合作经验是在介绍 se 课程, 他们修改开源项目在 6-8 组。然而, 学生们在这么大的团队中工作有问题, 觉得课程材料和项目"脱节"。我们决定在 2017 年重新设计这门课程, 试图在理论和实践以及技术和专业技能之间取得平衡, 每学期的最大课程工作量为 150 小时。我们在本文中分享了我们的经验, 讨论了我们用来改进团队合作和帮助学生以更自主的方式学习新技术的策略。我们还讨论了我们从两次教授新课程中学到的东西。少

2018 年 4 月 5 日提交;最初宣布 2018 年 4 月。

评论:8 页, 9 个表格, 4 个数字, 第二届千禧一代软件工程教育国际讲习班

msc 类: 97p70

70. 第 xiv:1804. 02822[[pdf](#), [ps](#),其他] Cs。马

多伊 10.1109/IWCIA.2015.7449480

基于网络增长模型的开源软件社区社会社区分析

作者:takumi ichimura, takuya uemoto

摘要: 开源 软件开发者的社会社区具有复杂的网络结构。网络结构表示项目与软件开发人员社区中的工程师之间的关系。一个项目形成了一些团队, 由被分成某一任务组的工程师组成。资料来源众所周知, 福格是开源网站之一。网络结构中的节点和弧线是指源锻造中的工程师及其在工程师之间的连接。在之前的研究中, 我们发现, 根据加入项目的开发者数量,项目的成长过程变得强劲。在成长阶段, 我们发现了代理商数量和生产项目数量之间的一些典型模式。通过这样的观察, 我们建立了一个执行项目成长过程的仿真模型。本文将陆军蚂蚁模型中所示的利他主义行为引入到软件开发人员的仿真模型中。通过实验仿真结果, 研究了软件开发过程的有效性。少

2018 年 4 月 9 日提交;最初宣布 2018 年 4 月。

评论:5 页, 5 个数字, ieee 第八届计算智能与应用国际研讨会 (iwcia2015) 的产品介绍

71. 第 xiv:1804. 02433[[pdf](#),其他] cse

野外可追溯性: 自动增强不完整的跟踪链接

作者:michael rath, jacob rendall, jin l. c.guo, jane Cleland-Huang 黄, patrick maeder

摘要: 软件和系统可追溯性被广泛接受为支持许多软件开发任务的基本要素。今天的版本控制系统提供内置功能, 允许开发人员使用一个或多个问题 id 标记每个提交, 从而提供构建块, 以便在功能请求、bug 之间建立项目范围内的可追溯性修复、提交、源代码和特定开发人员。然而, 我们对六个开源项目的分析显示, 平均只有 60% 的提交与具体问题有关。如果没有这些基本链接, 整个项目范围的链接集将是不完整的, 因此不值得信赖。在本文中, 我们讨论了提交和问题之间缺少链接的根本问题。我们的方法利用流程和文本相关功能的组合来描述问题和代码更改, 以训练分类器来识别提交消息中缺少的问题标记, 从而生成缺少的链接。我们进行了一系列实验, 以评估我们针对六个开源项目的方法, 并表明它能够有效地推荐链接, 以平均 96% 的召回率和 33% 的精度标记问题。在一项用于增加一组现有跟踪链接的相关任务中, 分类器在所有项目中的精度都高于 89%, 并召回 50% 较少

2018 年 4 月 6 日提交;最初宣布 2018 年 4 月。

评论:icse 2018

72. 第 [xiv:1804.02053](#)[pdf,其他] cse

指标仪表板: 软件质量度量的托管平台

作者:[george k. thiruvathukal](#), [shilpika](#), [nicolas j.hayward](#) , [konstantin läufer](#)

摘要: 科学界正在形成一种共识, 即科学研究的进展取决于 "各级软件的质量和可获得性" (wssspe.researchcomputing.org.uk/)。这一进展取决于采用软件工程中最好的传统--和新兴实践, 尤其是与更正式的软件工程传统交叉的敏捷实践。作为我们研究计算科学与工程 (cse) 软件开发项目过程中质量指标的更大探索性项目的第一步, 我们开发了指标仪表板, 这是一个用于生产和通过在 github 上挖掘开源软件存储库来观察指标。与 github 和提供单个性能指标 (例如提交) 的类似系统不同, 指标仪表板侧重于指示团队进度和项目运行状况的指标。指标仪表板允许用户提交托管存储库的 url 进行批处理分析, 然后对其结果进行缓存。完成后, 用户可以在数字和视觉上以交互方式研究随时间变化的各种指标 (以不同的粒度)。该系统的初始版本是作为公共云服务 (saas) 启动和运行的, 支持项目规模 (kloc)、缺陷密度、缺陷损坏和生产力。虽然我们的系统绝不是第一个支持软件指标的系统, 但我们相信它可能是任何托管项目都可以使用的第一个以社区为中心的可扩展资源之一。少

2018 年 4 月 8 日提交;v1 于 2018 年 4 月 5 日提交;最初宣布 2018 年 4 月。

评论:9 页

73. 第 [xiv:804.00832](#)[pdf,其他] Cs. CI

应序列到序列学习对语文本的音符性恢复

作者:[iroro orife](#)

摘要: Yorùbá是一种广泛使用的西非语言, 其书写系统具有丰富的声调和正字法。除了极少数例外, 由于设备和应用支持有限, 电子文本中省略了变音。对话提供了形态信息, 对于词汇消歧、发音至关重要, 对于任何 yorùbá文本到语音 (tts)、自动语音识别 (asr) 和自然语言处理 (nlp) 任务都是至关重要的。将自动分形恢复 (adr) 重构为机器翻译任务, 我们用两种不同的注意序列到序列神经模型对未变音文本进行了处理。在我们的评估数据集中, 这种方法产生的变光错误率小于 5%。我们发布了预先训练的模型、数据集和源代码, 作为一个开源项目, 以推进 yorùbá语言技术的发展。少

2018 年 10 月 29 日提交;v1 于 2018 年 4 月 3 日提交;最初宣布 2018 年 4 月。

评论:6 页, 3 个数字。讨论了 2018 年互动预印, 并发表了额外的数字和评审者的评论

74. 第 [xiv:1803.10344](#)[pdf,其他] 反渗透委员会

手持式控制无人机: 开源库

作者:[kathiravan natarajan](#), [trong-huy d. nguyen](#), [mutlu mete](#)

摘要: 无人机通常使用操纵杆、遥控器、移动应用程序和嵌入式计算机进行控制。这些方法的几个重要问题是, 无人机控制受到电磁辐射范围的限制, 容易受到干扰噪声的影响。在这项研究中, 我们建议使用手势作为控制无人机的方法。我们研究了计算机视觉方法的应用, 以开发一种直观的方式, 在无人机和其运营商之间进行无代理的通信。基于计算机视觉的方法依赖于无人机相机的能力来捕捉周围的图像, 并使用模式识别将图像转换为有意义的和可操作的信息。拟议的框架涉及采取最终行动的几个关键部分。它们是: 图像隔离从前置摄像头的视频流, 创建一个强大和可靠的图像识别基于隔

离的图像,最后将分类手势转换为可操作的无人机移动,如起飞,着陆,盘旋和等等。在这项工作中,研究了一组五个手势。基于 haar 特征的 adaboost 分类器用于手势识别。我们还设想操作人员的安全和无人机根据计算机视觉计算距离的行动。在考虑主要场景变异性、光照、背景和距离的情况下,进行了一系列的手势识别精度测量实验。分类精度显示,光线充足,背景清晰,在 3 英尺内的手势被正确识别超过 90%。还讨论了当前框架的局限性和更好地识别手势的可行解决方案。我们开发的软件库和手势数据集是在项目网站上开放的。少

2018 年 3 月 27 日提交;最初宣布 2018 年 3 月。

评论:icdis 2018

75. 第 1803. 10215[[pdf](#)] cs.PL

多伊 [10.2221/程正/](#)

深优先级冲突的零开销消歧

作者:luís eduardo de souza amorim, michael j.steindorfer, eelco visser

摘要: ** 上下文 ** 无上下文语法广泛用于语言原型设计和实现。它们允许以简洁和宣言的方式将特定于域或通用编程语言的语法形式化。然而,编写无上下文语法的自然而简洁的方法往往是模棱两可的。因此,语法形式主义支持以 * 声明性消歧规则 * 的形式进行扩展,以指定运算符优先级和关联性,从而解决由对应于表达式的语法子集引起的歧义。** 查询 ** 在解析器中实现对声明性消歧的支持通常伴随着以下一个或多个实践中的限制:缺乏解析性能或缺乏模块化(即不允许语法的构成可能不同的语言的片段)。后一个主题通常由无扫描体广义解析器处理。我们的目标是为无扫描语的广义解析器配备新的消歧方法,这些方法本质上是高性能的,而不会影响模块化和语言组成的问题。** 方法 ** 在本文中,我们提出了一种新的低开销实现技术,用于消除具有轻量级数据依赖关系的无扫描程序广义解析器中的深层关联性和优先级冲突的歧义。** 知识 ** 运算符优先级的歧义和关联性产生于组合一种语言的各种运算符。虽然 * 浅层冲突 * 可以通过一级树模式有效地解决,但 * 深层冲突 * 需要更复杂的技术,因为它们可以任意嵌套在树中。目前解决深度优先级冲突的最先进方法带来了严重的性能开销。** 接地 ** 我们根据最先进的声明性消歧机制评估了我们的新方法。通过解析用 java 和 ocaml 编写的流行开源存储库的语料库,我们发现,当解析具有深度优先级的程序时,我们的方法比语法重写技术产生高达 1.73 x 的速度。在没有深度冲突的情况下分析程序时,开销为 1-2%。** 重要性 ** 最近的一项实证研究表明,深刻的优先性冲突确实在现实世界的项目中广泛存在。研究表明,在 github 上的一系列流行的 ocaml 项目中,多达 17% 的源文件包含深度优先级冲突。但是,文献中没有解决深度优先级冲突的有效消除歧义问题的解决方案,并支持模块化和可组合的语法定义。少

2018 年 3 月 27 日提交;最初宣布 2018 年 3 月。

日记本参考:编程的艺术、科学和工程,2018,第 2 卷,第 3 期,第 13 条

76. 第 1803. 10198[[pdf](#)] cs.PL

多伊 [10.222/程报](#)

基于自动重构的新语言特征应用的主动经验评估--以 java 8 默认方法为例

作者:raffi k 孵化 adourian, hidehiko masuhara

摘要: 编程语言和平台会随着时间的推移而改进,有时会产生新的语言功能,从而提供许多好处。然而,尽管有这些好处,开发商可能并不总是愿意采用他们的项目,出于各

种原因。在本文中,我们描述了一个实证研究,我们评估了一个特定的新语言特征的采用。研究开发人员如何使用(或不使用)新的语言功能在编程语言研究和工程中非常重要,因为它使设计人员能够深入了解语言在该语言中创建意义程序的可用性。这些知识反过来又可以推动该领域未来的创新。在这里,我们将探讨 java 8 默认方法,它允许接口包含(实例)方法实现。默认方法可以简化界面演变,使某些无处不在的设计模式成为冗余,并提高模块化和可维护性。这项工作的重点是通过科学的方法和新的技术,发现开发人员认为这些构造有用的情况,而不是它们没有的情况,以及每一种情况的原因。尽管有几项研究的中心是评估新的语言特征,但就我们所知,这种结构以前没有被考虑过。尽管他们有好处,但我们发现开发人员并没有在所有情况下都采用默认方法。我们的研究包括提交拉请求,在不改变原始程序语义的情况下,向 19 个实际的开源 java 项目介绍语言功能。这种新颖的评估技术是主动的,因为采用的是自动重构方法,而不是等待开发人员发现和集成功能本身。通过这种方式,我们提出了更早而不是更晚有效使用语言功能的最佳实践和模式,并能够指导(近期)未来的语言进化。我们预见这种技术将有助于评估其他新的语言功能、设计模式和其他编程习语。少

2018 年 3 月 27 日提交;最初宣布 2018 年 3 月。

日记本参考:编程的艺术、科学和工程,2018,第 2 卷,第 3 期,第 6 条

77. 第 1803.10144[[pdf](#), [ps](#),其他] cse

多伊 [10.114/3393983196464](#)

公共 git 存档:适用于所有数据集的大代码数据集

作者:[vadim markovtsev](#), [waren long](#)

摘要:开源软件项目的数量呈指数级增长。主要的在线软件存储库主机 github 已经积累了数千万张可公开使用的 git 版本控制存储库。尽管现有的开放源代码所带来的研究潜力显然很大,但不存在重要的大规模开放源代码数据集。在本文中,我们提出了公共 git 存档--来自 github 的 1,2,014 个顶级书签 git 存储库的数据集。我们描述了新的数据检索管道来重现它。我们还详细介绍了执行数据集更新的策略和法律问题。公共 git 存档在磁盘上占用 3.0 tb,比当前源代码数据集大一个数量级。数据集通过 http 提供,并提供项目的源代码、相关元数据和开发历史记录。数据检索管道采用优化的工作人员队列模型和优化的存档格式,以有效地存储分叉的 git 存储库,从而减少下载和保留的数据量。公共 git 档案的目标是为“大代码”研究开辟无数新的机会。少

2018 年 3 月 20 日提交;最初宣布 2018 年 3 月。

78. 第 1803.10017[[pdf](#)] cs. cy

制造过程硬件和软件开发,以实施飞机制造设施管理的创新技术

作者:[s. e. pyatovsky](#), [a. n. serdyuchenko](#)

文摘:本文提出了基于管理决策实施创新技术的现代飞机制造设施竞争力增长的途径。本文结合俄罗斯经济对国际民用飞机制造商的依赖,建立了航空业与国民经济发展的联系。给出了俄罗斯和国际制造商生产的民用飞机的比较统计数据。对军事和民航进行了比较分析。研究表明,在一家高科技企业中实施基于开源和 olap 技术的项目是飞机制造设施竞争增长的先决条件。少

2018 年 3 月 27 日提交;最初宣布 2018 年 3 月。

日记本参考:经济与创业杂志 12 (2018) 1 (90) 856-859;<https://goo.gl/ZzZTaa>

79. 第 1803.09529[[pdf](#),其他] cse

多伊 10.114/333440.3394951

海报: 开源项目中的传播--电子邮件时代的终结?

作者: [verena käfer](#), [daniel graziotin](#), [ivan bogicevic](#), [stefan wagner](#), [jasmin ramadani](#)

文摘: 通信在软件工程中是必不可少的。特别是在分布式开源团队中, 通信需要得到包括邮件列表、论坛、问题跟踪器和聊天系统在内的渠道的支持。然而, 我们并不清楚开源项目中的利益攸关方使用哪些沟通渠道。在这项研究中, 我们通过调查 400 个 github 项目的统计代表性样本来填补知识空白。我们通过对项目数据的正则表达式发现了使用的通信通道。我们表明 (1) github 项目的一半使用可观察到的通信通道;(2) github 问题、电子邮件地址和现代聊天系统 gitter 是最常见的渠道;(3) 邮件列表只有五个, 并有一个低于所有现代聊天系统的总和市场份额。少

2018 年 3 月 26 日提交;最初宣布 2018 年 3 月。

评论 2 页。作为海报出现在第 40 届软件工程国际会议 (icse 2018) 的配套会议上

80. 第 xiv: 1803. 09166[[pdf](#),其他] [cs. cy](#)

"开始智能": 通过临床领域的方法, 对微创癌症治疗进行开放式、基于 web 的建模

作者: [phil weir](#), [roland ellerweg](#), [stephen payne](#), [dominic reuter](#), [tuomas alhonnoro](#), [phillipvogltreiter](#), [panchatcharam mariappan](#), [mika pollari](#), [chang sub 公园 peter voigt](#), [tim van Oostenbrugge](#), [sebastian fischer](#), [peter kalmar](#), [jürgen futterer](#), [ph 利普斯蒂格勒](#), [stepan zangos](#), [ronan flanagan michael moche](#), [marina kroesnik](#)

摘要: 临床医生通过场外无障碍、数据共享和专业互动, 受益于在线治疗规划系统。除了提高临床价值外, 仿真工具的集成还为开放式、多学科的研究合作提供了创新的途径。提出了一种可扩展的临床医生、技术人员、制造商和研究人员在仿真框架的基础上构建的系统。这是通过一个将理论、工程和临床领域的实体联系起来的域模型实现的, 它允许为几个开源求解器生成仿真配置的算法。该平台应用于微创癌症治疗 (mict), 允许介入放射科医生上传患者数据, 分割患者图像, 并验证射频消融、冷冻消融、微波消融和不可逆的电穿孔。传统的放射学软件布局在浏览器中提供临床使用, 具有简单的引导模拟, 主要用于培训和研究。开发人员和制造商访问基于 web 的系统来管理他们自己的模拟组件 (设备、数值模型和临床协议) 和相关参数。该系统由四个中心的介入放射科医生使用假名病人数据进行测试, 作为 "http://gosmart-project.eu 智能项目" 的一部分。仿真技术作为一组开源组件 <http://github.com/go-smart> 发布。少

2018 年 3 月 24 日提交;最初宣布 2018 年 3 月。

81. 第: 1803. 07856[[pdf](#)] [si](#)

多伊 10.114/3148330.3152699

在线社区内的协作调查: 软件开发与艺术创作

作者: [giuseppe iaffaldano](#)

摘要: 在线创意社区已经能够开发大型的开源软件 (oss) 项目, 如 linux 和 firefox, 在互联网上进行的成功合作。这些社区还扩展到动画、电子游戏和音乐等创意艺术领域。尽管它们越来越受欢迎, 但导致这些社区成功合作的因素并不完全清楚。下面, 我介绍了我的博士研究项目, 该项目旨在从从开放源代码软件社区文献中获得的经验出发, 改善创意艺术社区的沟通、协作和保留。少

2018 年 3 月 21 日提交;最初宣布 2018 年 3 月。

评论:2018 年, 博士学术讨论会, 2018 年 1 月 7-10 日, 美国佛罗里达州萨尼贝尔岛
类:H.5。3

日记本参考:2018 年支持集团工作会议论文集

82. 第 1803.06502[[pdf](#),其他] cse

基于马尔可夫链模型的开发人员情感分析

作者:[giuseppe Destefanis](#), [marco ortu](#), [steve 墓卖](#), [stephen swift](#), [roberto tonelli](#), [mic 其米歇尔·marchesi](#)

文摘: 在本文中, 我们对来自软件系统开源存储库的 500 多条注释进行了分析。我们的目的是根据经验确定开发者如何在开发者评论中表达的礼貌、情绪和情感所产生的某些心理条件下相互互动。参与开源项目的开发人员通常不认识对方; 他们主要通过邮件列表、聊天室和问题跟踪系统等工具进行沟通。他们的沟通方式影响项目所涉人员的发展过程和生产力。我们评估了开发人员发布的评论的礼貌、情绪和情绪, 并研究了沟通流程, 以了解他们在不礼貌和负面评论的情况下是如何互动的 (反之亦然)。我们的分析表明, 当出现不礼貌或负面评论时, 下一个评论不礼貌或否定的概率分别为 14% 和 25%;愤怒, 然而, 有 40% 的概率, 其次是进一步的愤怒评论。这一结果可以帮助管理者控制系统的开发阶段, 因为社会方面会严重影响开发人员的生产力。在分布式环境中, 这可能具有特定的共振。少

2018 年 3 月 17 日提交;最初宣布 2018 年 3 月。

83. 第 18005741[[pdf](#),其他] cse

多伊 [10.114/31919536.3 195848](#)

我们为什么要参与 floss: 来自核心开发人员的答案

作者:[jailton coelho](#), [marco tulio valente](#), [lu 西亚 a l.silva](#), [andre hora](#)

文摘: 自由自由开源 软件 (floss)项目的维护和发展要求核心开发人员不断发挥吸引力。在本文中, 我们报告了对 52 名开发者的调查结果, 他们最近成为流行的 github 项目的核心贡献者。我们揭示了他们在 floss 项目中扮演关键角色的动机 (例如, 改进项目, 因为他们也在使用这些项目), 这些项目的特点在他们的参与过程中帮助最大 (例如, 友好的社区), 以及接受调查的核心开发人员所面临的障碍 (例如, 项目负责人缺乏时间)。我们还将我们的结果与其他类型的开源 贡献者 (临时、一次性和新来者) 的相关研究进行了比较。少

2018 年 3 月 15 日提交;最初宣布 2018 年 3 月。

评论:2018 年 chase 会议接受: 第 11 届软件工程合作与人类方面国际研讨会 (8 页)

84. 第 xiv:18004818[[pdf](#),其他] Cs。Hc

智能用户界面深度学习工具包与库的研究

作者:[jan zacharias](#), [michael barz](#), [daniel sonntag](#)

摘要: 本文概述了著名的深度学习工具包, 特别是最近的出版物, 这些出版物为实现智能用户界面 (iui) 中常见的任务提供了开源软件。我们为计划在 iui 研发项目中使用深度学习技术的研究人员和软件工程师提供科学参考。少

2018 年 3 月 14 日提交;v1 于 2018 年 3 月 13 日提交;最初宣布 2018 年 3 月。

类:H.5。2

85. 第 1803.04129[[pdf](#),其他] cse

捐赠徽章有吸引力吗? 开发人员对 eclipse bug 报告响应的案例研究

作者: [中泽凯太郎](#), [hata hideaki](#) [hata](#), [kenichimatsumoto](#)

摘要: eclipse 是一个开源软件项目, 它通过在其问题跟踪系统 bugzilla 中展示捐赠徽章来承认其捐助者。然而, 这一战略的奖励效果目前尚不清楚。我们应用因果推断的框架来调查开发人员对带有捐赠徽章的 bug 报告的响应相对及时性与没有徽章的 bug 报告相比相对较快, 并估计捐赠徽章会将开发人员响应时间缩短中值时间约两小时。捐赠徽章的出现对捐助者和组织者都很有吸引力, 因为它具有实际、有益但廉价的效果。少

2018 年 7 月 18 日提交;v1 于 2018 年 3 月 12 日提交;最初宣布 2018 年 3 月。

评论:4 页, ieee 软件

86. 第 xiv:1803.03784[[pdf](#), [ps](#), [其他](#)] 反渗透委员会

任务约束轨迹优化的交替投影与增强拉格朗日相结合的方法

作者: [arun kumar singh](#), [reza Ghabcheloo](#), [andreas muller](#), [harit pandya](#)

摘要: 在任务空间约束下, 机械手的运动规划是困难的, 因为它限制了关节配置始终位于隐式定义的流形上。可以将任务约束运动规划看作是一个具有非线性等约束的优化问题, 一般的非线性优化技术可以解决。在本文中, 我们提出了一种新的自定义优化器, 它利用了许多任务约束中存在的底层结构。我们的方法的核心是一些简单的重新配方, 当与交替投影的强调方法结合使用时, 可以实现基于凸优化的高效例程, 从而计算任务约束的可行解决方案。随后, 我们在这一结果的基础上, 使用增强拉格朗日的概念来指导可行的解决方案, 以实现那些同时最大限度地减少用户定义的成本函数的解决方案。我们证明了所提出的优化器是完全分布式的, 因此可以很容易地并行化。我们验证了一些常见的机器人基准问题的配方。特别是, 我们证明了所提出的优化器在与任务空间中类似的自然轨迹相对应的关节空间中实现了循环运动。此外, 作为基线, 我们将建议的优化器与开源包 s 蒂中提供的现成非线性求解器进行比较。我们表明, 对于类似的任务约束残差和平滑成本, 它可以比 sicy 替代方案快三倍以上。少

2018 年 3 月 10 日提交;最初宣布 2018 年 3 月。

评论:8 页

87. 第 180344 次[[pdf](#)] Cs. 简历

利用神经网络建模和张力流解决傅立叶 ptychoms 形图的问题

作者: [蒋少伟](#), [郭凯凯](#), [廖军](#), [郑国安](#)

摘要: 傅立叶 ptychotr 是一种最近开发的用于大视场和高分辨率显微镜的成像方法。在这里, 我们利用卷积神经网络 (cnn) 对傅立叶 ptychen-sprisri-pjspri-prispri-pjesultspri-pjesulte 的成像过程进行了建模, 并对复杂的目标信息进行了恢复。在这种方法中, 网络的输入是空间域中的点扩散函数或傅里叶域中的相干传递函数。该对象被视为卷积或乘法层的二维可学习权重。网络的输出被建模为我们旨在最小化的损耗函数。网络的批处理大小对应于一个向前/向后传递中捕获的低分辨率图像的数量。我们使用一个流行的开源机器学习库 tensorflow 来建立网络和进行优化过程。我们分析不同学习速率、不同求解器和不同批次大小的性能。结果表明, 使用 adam 优化器的大批处理尺寸通常可实现最佳性能。为了加快相位检索过程, 我们还讨论了一种利用乘法神经网络模型实现傅里叶变换级投影的策略。由于卷积和乘法是成像建模中最常见的两种运算, 因此所报告的方法可以为检验许多相干和非相干系统提供一个新的视角。作为演示, 我们讨论了用于建模单像素成像和结构化

照明显微镜 (sim) 的所报告网络的扩展。使用 sim 的神经网络演示了 4 帧分辨率加倍。我们已经为广泛的研究界开放了我们的实施代码。少

2018 年 3 月 9 日提交;最初宣布 2018 年 3 月。

88. 建议: 1803.03175[[pdf](#)] cse

多伊 [10.18293/SEKE2018-085](#)

大型开源生态系统公共发展项目的自动检测: 基于 github 的探索性研究

作者: [易成](#), [李兵](#), [李增阳](#), [彭亮](#)

摘要: github 托管了超过 1000 万个软件项目, 是研究开发人员和软件项目行为的最重要数据源之一。但是, 随着开源数据集大小的增加, 挖掘这些数据集的潜在威胁也在增加。随着数据集的增长, 人类确认所有样本的质量变得逐渐不现实。一些研究调查了这一问题, 并提供了避免样本选择威胁的解决方案, 但其中一些解决方案 (例如, 寻找发展项目) 需要人为干预。当需要处理的数据量增加时, 这些半自动解决方案就变得不那么有用了, 因为需要人为干预的努力远远超出了负担得起的范围。为了解决这一问题, 我们对 ghtorrent 数据集进行了调查, 并提出了一种公共开发项目的检测方法。结果表明, 该方法能通过两种方式有效地改进样品选择过程: (1) 提供一个简单的样品自动选择模型 (精度为 0.827, 召回 0.827); (2) 我们还提供了一个复杂的模型, 以帮助研究人员仔细筛选样品 (比手动确认所有样品的工作量减少了 63.2, 并可实现 0.926 精度和 0.926 召回)。少

2018 年 5 月 8 日提交;v1 于 2018 年 3 月 8 日提交;最初宣布 2018 年 3 月。

评论: 被 seke2018 会议接受

89. 第 xiv: 18002205[[pdf](#), [ps](#), [其他](#)] cse

多伊 [10.114/33399.3183411](#)

代码审阅注释: 语言问题

作者: [vasiliki efstathiou](#), [diomidis spinellis](#)

摘要: 最近的研究证明, 协作软件开发中的有效通信对软件开发生命周期有重大影响。尽管相关的定性和定量研究指出了格式良好的信息的文本特征, 但交织在一起的语言结构的基本语义在很大程度上仍然被误解或忽视。特别是关于代码评审的质量, 彻底反馈的重要性, 经常提到明确的理由, 但很少与相关的语言特征联系在一起。作为解决这一缺陷的第一步, 我们建议将这些研究建立在语言学理论的基础上。我们特别关注连贯语言的语言结构, 并解释如何在实践中加以利用。我们对相关方法进行了反思, 并通过四个开源项目的初步研究, 研究了现有发现之间可能存在的联系, 以及我们建议的检测有用代码的文本特征的方向评论。少

2018 年 3 月 6 日提交;最初宣布 2018 年 3 月。

90. 第 xiv: 18001126[[pdf](#)] cse

多伊 [10.114/336369.3196372](#)

大型分布式项目中的开发商人格--以阿帕奇生态系统为例

作者: [fabio calefato](#), [giuseppe iaffaldano](#), [Filippo lanubile](#), [bogdan vasilescu](#)

摘要: 大型分布式项目通常是多个开发人员集体努力的结果, 每个开发人员都有不同的个性。对开发者个性的研究有可能解释他们在各种背景下的行为。例如, 信任他人的倾向----全球软件工程成功的一个关键因素----被认为对分布式项目中的代码评审结果产生了积极影响。在本文中, 我们对开源软件项目中开发人员的个性进行了定量分析,

旨在作为分布式项目的一种极端形式,在这种形式中,没有一个组织控制项目中。我们从从事 apache 软件基金会 (asf) 项目的开发人员提供的代码提交和电子邮件中挖掘生态系统级数据,作为大型规模分布项目的代表。我们发现,随着时间的推移,开发者变得更加认真、愉快和神经质。此外,个性特征并不因其作用、成员和对项目的贡献程度而有所不同。我们还发现了更开放、更令人愉快的开发者更有可能成为项目贡献者的证据。少

2018 年 9 月 24 日提交;v1 于 2018 年 3 月 3 日提交;最初宣布 2018 年 3 月。

评论:2018 年 5 月 28 日至 29 日,瑞典哥德堡,《全球软件工程国际方案》(icgseo18)

91. 建议: 1802. 10233[[pdf](#),[其他](#)] Cs. Db

apache 方解石: 对异构数据源进行优化查询处理的基本框架

作者:[edmon begoli](#), [jesús camacho rodríguez](#), [julian hyde](#), [michael j. mior](#), [daniel lemire](#)

摘要: apache 方解石是一个基础软件框架,它为许多流行的开源数据处理系统提供查询处理、优化和查询语言支持,如 apache hive、apache 风暴、apache flink、druid、和地图。方解石的体系结构包括一个模块化和可扩展的查询优化器,其中包含数百种内置优化规则、一个能够处理各种查询语言的查询处理器、一个为可扩展性而设计的适配器体系结构以及对异构数据模型和存储(关系、半结构化、流媒体和地理空间)。这种灵活、可嵌入和可扩展的体系结构使方解石成为大数据框架中采用的极具吸引力的选择。它是一个活动项目,继续引入对新型数据源、查询语言以及查询处理和优化方法的支持。少

2018 年 2 月 27 日提交;最初宣布 2018 年 2 月。

评论:simoda18

92. 建议: 1802. 09335[[pdf](#),[其他](#)] cs. cy

城市土地利用、旅游需求和交通分配建模的综合管道体系结构

作者:[paul waddell](#), [geoff](#) 波音公司, [max gardner](#), [emily porter](#)

摘要: 整合土地使用、旅行需求和交通模式是区域规划的黄金标准,但很少以有意义的方式实现,特别是在分类数据的规模上。在本报告中,我们提出了一个新的管道架构,用于城市土地利用、出行需求和交通分配的综合建模。我们的土地利用模型 urbansim 是一个开源的微仿真平台,被世界各地的都市规划组织用于模拟城市在漫长(~ 30 年)时间范围内的增长和发展。urbansim 作为情景分析工具特别强大,使规划人员能够以统计严格的方式比较和对比不同政策决定对长期土地使用预测的影响。我们的旅行需求模型 activitysim 是一个基于代理的建模平台,可生成合成的原始目的地旅行需求数据。最后,我们使用基于 frank-wolfe 算法的静态用户均衡流量分配模型将车辆分配到特定的网络路径,以便在起源和目的地之间进行旅行。此流量分配模型在高性能计算环境中运行。由此产生的拥堵行驶时间数据可以反馈到 urbansim 和 activitysim 进行下一次模型运行。本技术报告介绍了这一研究领域,介绍了该项目迄今在开发这一综合管道方面取得的成就,并提出了即将出台的研究议程。少

2018 年 2 月 7 日提交;最初宣布 2018 年 2 月。

93. 建议: 1802.08547[[pdf](#),[其他](#)] cse

多伊 [10.11145/3183519.3183554](#)

智能单元: 工业嵌入式软件自动化单元测试的经验评估

作者:张成宇,严一晨,周汉鲁,姚银波,克武,苏婷,苗伟凯,普格光

摘要: 本文针对嵌入式软件的自动化单元覆盖测试。为了实现这一目标,通过分析工业需求和我们以前在自动化单元测试工具 `caut` 方面的工作,我们重建了一个新的工具,智能单元,以解决我们的合作伙伴公司中的工程需求。`smartunit` 是一个动态符号执行实现,它支持语句、分支、边界值和 `mc/dc` 覆盖率。`smartunit` 已被用来在实际项目中测试 100 多万行代码。出于保密动机,我们选择三个内部真实项目进行实证评估。我们还对两个开源数据库项目 `sqlite` 和 `postgresql` 进行评估,以测试我们的工具的可伸缩性,因为嵌入式软件项目的规模大多不大,5k-50k 行平均代码。从我们的实验结果,一般来说,90% 以上的业务功能在商业嵌入式软件中实现 100% 语句、分支、`mc/dc` 覆盖率,80% 以上的 `sqlite` 函数实现 100% `mc/dc` 覆盖率,在 `postgresql` 中实现 60% 以上的函数实现 100% 的 `mc/dc` 覆盖。此外,`smartunit` 能够在单元测试级别找到运行时异常。我们还报告了诸如超出边界的数组索引和 `sqlite` 中的零除等异常。此外,我们还分析了自动化单元测试在我们的环境中覆盖率低的原因,并对工业中自动化单元测试的手动单元测试情况进行了调查。少

2018 年 6 月 17 日提交;v1 于 2018 年 2 月 23 日提交;最初宣布 2018 年 2 月。

评论:在《第 40 届软件工程国际会议论文集:软件工程实践中的轨道》,2018 年 5 月 27 日至 3 日 (icse-seip '18), 10 页

94. [建议: 1802.08522\[pdf,其他\]](#) cs. it

多伊 10.1049/joe.2014.0055

simcommsys: 消除纠错代码模拟中的错误

作者:john a. briffa, stepfan wweemeyer

文摘: 在本文中,我们介绍了模拟,一个通信系统的模拟器,我们正在开源许可证下发布。该项目的核心是一组定义通信系统组件的 `c++` 库和一个分布式蒙特卡洛模拟器。主要感兴趣的是错误控制编码组件,在该组件中实现了各种二进制和非二进制代码,包括涡轮增压、ldpc、废除积累和 reed-solomon。该项目还包含许多实现通信系统各个阶段(如编码器和解码器)的现成生成二进制文件、一个完整的模拟器和一个系统基准。最后,simcommsys 还提供了许多 shell 和 python 脚本来封装常规用例。只要所需的组件在 simcommsys 中已经可用,用户就可以模拟自己设计的完整通信系统,而无需进行任何额外的编程。开发(只需要实现新组件)和使用(模拟特定结构)的严格分离鼓励了实验工作的重现性,并降低了出错的可能性。在对框架进行概述之后,我们提供了一些如何使用该框架的示例,包括简单编解码器的实现、通信系统的规范及其模拟。少

2018 年 2 月 23 日提交;最初宣布 2018 年 2 月。

日记本参考:j. a. briffa 和 s. wsemeyer, "Simcommsys: 从纠错代码模拟中消除错误", iet 工程杂志, 2014 年 6 月

95. [xiv:1802.06025\[pdf,其他\]](#) cse

元学习在跨项目缺陷预测中的系统研究

作者:费米森波尔图,莱安德罗·明库,埃米莉亚·门德斯,阿德尼索·西芒

文摘: 基于外部项目数据的目标项目缺陷预测称为跨项目缺陷预测 (cpdp)。提出了几种提高 cpdp 模型预测性能的方法。然而,最先进的方法之间缺乏比较。此外,以前的工作表明,最适合项目的方法可以根据所预测的项目而有所不同。这使得选择使用哪种方法变得困难。我们对从最先进的方法中得出的 31 种 cpdp 方法进行了广泛的实验比较,

这些方法适用于 15 个开源软件项目的 47 个版本。有四种方法突出显示了跨数据集的最佳性能。然而, 这些方法中最适合的方法仍然因预测的项目而异。因此, 我们提出并评估了一种元学习解决方案, 该解决方案旨在自动选择和推荐最适合项目的 cpdp 方法。我们的研究表明, 元学习解决方案能够从以往的经验中学习, 动态推荐合适的方法。然而, 与基本方法相比, 所提出的解决方案在性能上差别不大。这些结果提供了关于应用于 cpdp 的元学习解决方案的可能性和局限性的宝贵知识。少

2018 年 2 月 16 日提交;最初宣布 2018 年 2 月。

评论:2017 年 11 月 6 日首次提交 [ieee](#) 软件工程交易

96. [建议: 180005084](#)[pdf,其他] [cse](#)

[多伊](#) [10.114/3180155.3180193](#)

程序员是在晚上工作还是在周末工作?

作者:[maëlick claes](#), [mika mäntylä](#), [miikka kuutila](#), [bram adams](#)

摘要: 不正常的工作时间会降低工作健康、总体福祉和生产力, 独立于某一职业。为今后的自动应力和过载检测方法提供参考, 本文建立了软件工程师工作模式的经验收集方法。为此, 我们通过调查 86 个大型开源软件项目(包括雇佣项目和志愿者) 的投入活动的时间戳, 对软件工程师的工作时间进行了首次大规模研究开发 人员。我们发现, 三分之二的软件工程师主要遵循典型的办公时间, 根据经验确定为从 10 小时到 18h, 通常不在夜间和周末工作。项目和个人之间存在很大差异。令人惊讶的是, 我们没有发现项目成熟会减少异常工作时间的支持。在 firefox 案例研究中, 我们发现雇佣的开发人员在办公时间内工作更多, 而资历, 无论是提交数量还是工作状态, 都不会影响工作时间。我们的结论是, 使用工作时间或工作产品的时间戳来检测压力, 需要在个人一级确定基线。少

2018 年 2 月 14 日提交;最初宣布 2018 年 2 月。

期刊参考:第四十届软件工程国际会议, 2018

97. [建议: 1802.03589](#)[pdf] [Cs](#)。直流

[多伊](#) [10.17559/电视-20141006121105](#)

使用 **mapreduce** 在云上进行分布式日志分析

作者:[galip aydin](#), [ibrahim riza hallac](#)

摘要: 本文介绍了基于部署在小云上的流行 mapreduce 框架的基于 web 的分布式数据分析系统的设计工作;专为分析 web 服务器日志而开发的。日志分析系统由多个群集节点组成, 它将分布式文件系统上的大型日志文件拆分为多个节点, 并使用 mapreduce 编程模型快速处理这些文件。群集是使用开源云基础架构创建的, 这使我们能够通过添加新节点轻松扩展计算能力。这使我们能够根据数据分析要求自动调整群集的大小。我们实施了 mapreduce 程序, 用于基本的日志分析需求, 如频率分析、错误检测、繁忙时间检测等, 以及需要运行多个作业的更复杂的分析。该系统可以自动识别和分析几种 web 服务器日志类型, 如 apache、iis、squid 等。我们使用开源项目来创建云基础架构和运行 mapreduce 作业。少

2018 年 2 月 10 日提交;最初宣布 2018 年 2 月。

98. [建议: 1802.03422](#)[pdf,其他] [cse](#)

gis 软件的实践现状

作者:[w. spencer smith](#), [adam Lazzarato](#), [jacques carette](#)

文摘: 我们提出了一种可重现的方法来分析特定科学领域的软件开发实践状态, 并将该方法应用于地理信息系统 (gis)。分析的基础是使用基于 13 个软件质量的 56 个问题模板对一组 30 个 gis 产品进行分级。产品的范围和用途从完整的桌面 gis 系统到独立的工具, 再到编程库包。产品的最终排名是使用层次分析法 (ahp) 确定的, 层次分析法 (ahp) 是一种多准则决策方法, 侧重于产品之间的相对比较, 而不是直接测量质量。研究结果揭示了一些 gis 软件的正确性、可维护性、透明度和重现性。三个建议作为反馈提供给 gis 社区: 1) 确保每个项目都有需求规范文档, 2) 提供丰富的支持方法, 如 irc (互联网中继聊天) 通道, 新的堆栈 exchange 标签问题, 或打开支持请求的问题跟踪器, 以及更传统的基于电子邮件的方法, 和, 3) 设计产品网站以实现最大的透明度 (开发过程), 以实现 开源项目, 提供开发人员指南。少

2018 年 2 月 9 日提交;最初宣布 2018 年 2 月。

99. **建议: 1802.03142**[pdf, ps,其他] Cs. CI

法语翻译对图书馆的扩充: 直接语音翻译评价的多模态语料库

作者:ali can Kocabiyikoglu, laurent besacier, olivier klaif

摘要: 最近在口语翻译 (slt) 方面的工作试图建立端到端的语音到文本翻译, 而不使用源语言转录在学习或解码过程中。然而, 虽然大量的并行文本 (如 europarl、open 字幕) 可用于培训机器翻译系统, 但没有大型 (100h) 和 开源 并行语料库包括源中的语音与目标语言中的文本对齐的语言。本文试图通过增加现有的 (单语) 语料库--"自由语言" 来填补这一空白。这个用于自动语音识别的语料库来自于 librivox 项目中的有声读物, 并已被仔细分割和对齐。在收集了与 libri 语音的英文电子书相对应的法语电子书后, 我们将句子级别的语音片段与它们各自的翻译进行了调整, 并获得了 236h 的可用并行数据。本文介绍了处理的细节, 以及对语料库的一个小子集进行的人工评价。该评价表明, 自动对齐得分与人类对双语对齐质量的判断有较好的相关性。我们相信, 这个语料库 (可在网上提供) 是有用的可复制的实验, 在直接语音翻译或更一般的口语翻译实验。少

2018 年 2 月 9 日提交;最初宣布 2018 年 2 月。

评论:lrec 2018, 日本

100. **建议: 1802. 01168**[pdf] Cs. DI

机器学习与规则和开箱即用与再培训: 开源参考和引文分析器的评估

作者:dominika tkaczyk, andrew collins, paric sheridan, joeran beel

摘要: 书目参考分析是指从书目参考字符串中提取机器可读的元数据, 如作者的姓名、标题或期刊名称。到目前为止, 已经提出了许多解决这一问题的方法, 包括正则表达式、知识库和监督机器学习。还提供了许多基于各种算法的 开源 参考解析器。在本文中, 我们应用、评估和比较了特定业务用例中的十个参考解析工具。这些工具是 anystyle-parter、Biblio、cermine、引文、cit-parter、grobid、parcit、pdfssa4met、参考匕首和科学分析, 我们在它们的开箱即用版本和根据特定项目数据进行比较。根据我们的评估, 最好的现成工具是得出来的工具是得出来的 (f1 0.89), 其次是 cermine (f1 0.89) 和 parscit (f1 0.89)。我们还发现, 尽管基于规则或正则表达式的基于机器学习的工具和工具平均实现了相似的精度 (基于 ml 的工具为 0.77, 而非基于 ml 的工具为 0.77), 但应用基于机器学习的工具会导致召回比基于 ml 的工具高出三倍 (0.66 对 0.66)。我们的研究还证实, 调整模型以特定任务的数据可以提高质量。在所有情况下, 重新训练的引用解析器版本都优于开箱即用的解析器:grobid f1 增加 3% (0.92

对 0.92), 由 c 原定 e 增加 11% (0.92 对 0.92), 为 parscit 增加 16% (0.92 对 0.92)。少

2018 年 4 月 19 日提交;v1 于 2018 年 2 月 4 日提交;最初宣布 2018 年 2 月。

评论:将在 2018 年数字图书馆联席会议会议上发表

101. 建议: 1802. 00921[[pdf](#),其他] cse

一种基于树的软件缺陷预测深模型

作者:[hoa khanh dam](#), [trang pham](#), [shienwee ng](#), [truyen tran](#), [john grundy](#), [aditya ghose](#), [taeksu kim](#), [chul-joo kim](#)

摘要: 缺陷在软件系统中很常见, 可能会给软件用户带来各种问题。已经开发了不同的方法来快速预测大型代码库中最可能出现的缺陷位置。它们中的大多数侧重于设计与可能有缺陷的代码相关的功能 (例如复杂性度量)。然而, 这些方法并不能充分捕获源代码的语法和不同级别的语义, 这是构建准确预测模型的重要功能。本文开发了一种新的预测模型, 该模型能够自动学习表示源代码的特征, 并将其用于缺陷预测。我们的预测系统建立在强大的深度学习, 树状结构的长期记忆网络, 直接匹配的抽象语法树表示的源代码。对两个数据集进行了评估, 一个来自三星提供的[开源项目](#), 另一个来自公共承诺存储库, 这证明了我们的项目内和项目方法的有效性。[跨项目预测](#)。少

2018 年 2 月 3 日提交;最初宣布 2018 年 2 月。

102. [xiv:1801. 10228](#)[[pdf](#),其他] Cs. 直流

[多伊](#) [10.114/3190508.3190538](#)

超类帐结构: 一种适用于已有存储的区块链的分布式操作系统

作者:[elli androulaki](#), [artem 谈判者](#), [vita bortnikov](#), [christian cachin](#), [konstantinos christidis](#), [angelo de caro](#), [david enyeart](#), [christopher fer 利](#)
[斯](#), [gennady laventman](#), [yacov manevich](#), [srinivasan muralidharan](#), [chetmurthy](#), [binh nguyen](#), [manishsethi](#), [gari singh](#), [keith smith](#), [alessandro somiotti](#), [chrysoula stathakopoulou](#), [marko vukolić](#), [sharon weed cocco](#), [jason yellick](#)

摘要: 结构是一个模块化和可扩展的开源系统, 用于部署和操作允许的区块链, 也是 linux 基金会 ([www.hyperledger.org](#)) 托管的超分类帐项目之一。结构是第一个真正可扩展的用于运行分布式应用程序的区块链系统。它支持模块化的协商一致协议, 使系统能够针对特定的用例和信任模型进行定制。结构也是第一个运行以标准通用编程语言编写的分布式应用程序的区块链系统, 而不需要对本机加密货币进行系统依赖。这与现有的区块链平台形成鲜明对比, 这些平台要求 "智能合同" 以特定于域的语言编写, 或者依赖于加密货币。结构实现了使用可移植的成员身份概念的允许模型, 该概念可以与行业标准身份管理集成。为了支持这种灵活性, fabric 引入了一种全新的区块链设计, 并改进了区块链应对不确定性、资源耗尽和性能攻击的方式。本文介绍了 fabric、其体系结构、各种设计决策背后的基本原理、最突出的实现方面以及分布式应用程序编程模型。我们通过实施比特币启发的数字货币并对其进行基准测试, 进一步评估结构。我们展示了 fabric 在某些流行的部署配置中实现了每秒超过 3500 事务的端到端吞吐量, 具有亚秒延迟, 并可很好地扩展到 100 多个对等方。少

2018 年 4 月 17 日提交;v1 于 2018 年 1 月 30 日提交;最初宣布 2018 年 1 月。

评论:出现在欧洲之家 2018 年会议的议事过程中

103. 建议: 1801. 09847[[pdf](#),其他] Cs. 简历

open3d: 用于 3d 数据处理的现代库

作者:周千一,杰西克公园,弗拉德伦·科尔通

摘要: open3d 是一个开源库,支持处理 3d 数据的软件的快速发展.open3d 前端在 c++ 和 python 中都公开了一组精心挑选的数据结构和算法。后端经过高度优化,并设置为并行化。open3d 是从一个干净的石板开发的,它有一组经过仔细考虑的小依赖关系。它可以在不同的平台上设置,并以最小的努力从源编译。代码是干净的、一致的样式,并通过清晰的代码审查机制进行维护。open3d 已在许多已发布的研究项目中使用,并被积极部署在云中。我们欢迎开源社区的贡献。少

2018 年 1 月 29 日提交;最初宣布 2018 年 1 月。

评论:<http://www.open3d.org>

104. 新建: 1801. 06268[pdf,其他] cse

开发人员如何测试 android 应用程序?

作者:mario linaires vasquez, carlos bernal-cardenas, kevin moran , denys poshyvanyk

摘要: 实现移动应用程序的完全自动化测试最近已成为研究人员和从业人员的一个重要研究课题。已经提出了大量的工具和方法来帮助移动开发人员,既可以增加手动测试实践,也可以自动化测试过程的各个部分。但是,目前的自动化测试方法不能让开发人员相信它们的好处,导致大多数移动测试都是手动执行的。为了帮助研究人员和从业者--他们设计了支持移动测试的方法--了解开发者的需求,我们分析了 102 位 android 项目开源贡献者的调查回复在执行测试时了解他们的做法。调查的重点是有关开发人员测试人员的做法和偏好的问题,以便 (i) 设计和生成测试用例,(ii) 自动化测试实践,以及 (iii) 对质量指标的感知,例如用于确定代码覆盖率的代码覆盖率。测试质量。通过分析从这次调查中收集到的信息,我们汇编了大量知识,以帮助指导研究人员和专业人员定制新的自动化测试方法,满足各种开源开发人员的需求。少

2018 年 1 月 18 日提交;最初宣布 2018 年 1 月。

评论:11 页,被第 33 届 ieee 软件维护和进化国际会议论文集 (icsmee17) 接受

105. 建议: 1801. 06 144[pdf,其他] cs.PL

一种用于指定和实现正确的脉冲流模块的形式化

作者:erick lavoie, la 里 ie hendren

摘要: pul-以便是一种基于回调函数的 javascript 请求驱动的功能设计模式,它可以创建和轻松组合用于创建流应用程序的独立模块。它被用于流行的开源项目,它周围的社区已经创建了一百多个兼容的模块。虽然对拉流设计模式的描述看起来很简单,但它确实表现出复杂的终止案例。尽管拉流设计模式很受欢迎,而且被广泛采用,但目前还没有任何正式规范可以帮助程序员对其实现的正确性进行推理。因此,本文的主要贡献是提供一个形式化的指定和实现正确的拉流模块的基础上:(1) 我们表明,拉流设计模式是声明性并发编程的一种形式;(2) 我们提出了一种基于事件的协议语言,它支持我们的形式化,独立于 javascript;(3) 我们对在两个模块的接口上发生的事件的预期序列提供了第一个精确和明确的定义,我们称之为拉流协议;(4) 我们指定了显示拉流协议的全部行为的参考模块;(5) 我们根据社区的期望验证我们的定义,方法是针对它们测试现有的核心拉流模块,并确定现有模块中未指定的行为。我们的方法有助于更好地理解拉流协议,确保社区模块的互操作性,并在论文和文档中简洁、精确地指定新的拉流抽象。少

2018 年 1 月 18 日提交;最初宣布 2018 年 1 月。

类:d.3 米

106. [建议: 1801 1.05932\[pdf,其他\]](#) cse

增强移动应用的 bug 报告

作者:[凯文·莫兰](#)

摘要: 现代软件开发领域的重点已经转向移动应用, 将其作为几乎无处不在的 "智能" 设备。由于这一趋势, 移动应用程序的复杂性一直在增加, 因此开发和维护尤其具有挑战性。但是, 很明显, 当前的 bug 跟踪系统无法有效地支持使用可操作信息构建报告, 从而直接导致错误的解决。为了满足改进报告系统的需要, 我们引入了一种名为 flusion 的新解决方案, 该解决方案可帮助用户自动完成移动应用错误报告中的复制步骤。博物馆将用户提供的信息链接到通过测试或发布前执行的静态和动态分析提取的程序工件。flusion 采用的方法可推广到当前的其他移动软件平台, 是移动软件项目进行设备外错误报告的一种新方法。我们通过进行一项研究来评估 fusion, 该研究定量和定性地衡量了系统的用户体验, 用于报告和复制 bug, 以及它所产生的 bug 报告的质量。在一项涉及 28 名参与者的研究中, 我们应用混淆来支持在 14 个开源 android 应用中发现的 15 个现实世界中的缺陷的报告和重现缺陷的维护任务。我们的研究结果表明, 与传统的 bug 跟踪系统相比, "融合" 可以帮助用户报告更详细的特定于应用程序的信息, 从而更可靠地复制报告中的 bug。少

2018 年 1 月 17 日提交;最初宣布 2018 年 1 月。

评论:77 页, 硕士论文提交给教师 @ 威廉玛丽学院

107. [第 1712 09152\[pdf\]](#) Cs。直流

以 br 高达 stack 为例部署高度安全的 openstack 云基础架构的技术报告

作者:[bashir mohammed](#), [sibuso moyo](#), [k. m maiyama](#), [sulayman kinteh](#), [al noaman m. k. Al-Shaidy](#), [m. a. kamala](#), m. kiran

摘要: 云计算已经成为一种流行的范式和提供可靠的分布式计算模型的有吸引力的模型, 它越来越受到学术研究和工业活动的极大关注。对于各种规模的机构和组织来说, 云部署是最重要的。一个灵活、免费的开源云平台的可用性是至关重要的, 该平台设计时没有适当的软件, 并能够与遗留系统和第三方应用程序集成。开放堆栈是根据 apache 许可证条款发布的免费开源软件, 具有分散和分布式的体系结构, 使其具有高度的灵活性。该项目启动后, 旨在设计一个名为 bradstack 的安全云基础设施, 该基础架构是建立在布拉德福德大学计算实验室的 openstack 上的。在本报告中, 我们介绍并讨论了部署安全的 br 高达多节点云基础架构和在 openstack 服务上进行渗透测试以验证 brandstack 平台上安全控制的有效性所需的步骤。本报告是一个切实可行的指导方针, 侧重于安全和与安全和实际基础设施有关的问题。它还为研究实施安全云解决方案的可能性的机构提供了参考。少

2017 年 12 月 25 日提交;最初宣布 2017 年 12 月。

评论:38 页 19 个数字