软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(6):1327−1329 [doi: 10.13328/j.cnki.jos.005233]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

大数据时代软件工程专题前言*

刘 璘[1], 周明辉[2], 尹 刚[3]

[1](清华大学 软件学院,北京 100084)

[2](北京大学 信息科学技术学院,北京 100871)

[3](国防科学技术大学 计算机学院,湖南 长沙 410073)

通讯作者: 刘璘, E-mail: linliu@tsinghua.edu.cn

中文引用格式: 刘璘,周明辉,尹刚.大数据时代软件工程专题前言.软件学报,2017,28(6):1327−1329. http://www.jos.org.cn/1000-9825/5233.htm

软件在人类社会生活中发挥着越来越重要的作用,软件工程研究软件系统构造、开发、运行、维护、演化的创新方法以提高效率和质量.从 20 世纪五、六十年代起,软件工程经历了从结构化到面向对象、网络服务化的演进.软件工程研究内容和范围不断扩展,其应用和实践也不断延伸,正发展成为结合人工智能、社会计算、认知科学、数据科学与工程等多学科交叉的领域.

为及时反映我国学者在结合大数据技术的软件工程研究新进展以及在大数据软件应用开发方面的最新实践经验,“大数据时代的软件工程”专题围绕上述新兴热点问题,同时也兼顾经典问题的最新突破,征集本领域近期取得的原创性研究成果.专题的征文范围包括(但不限于)面向特定领域的大数据应用开发、调试、部署及运行管理过程中的软件工程问题、软件工程数据质量问题、面向开源软件生态系统的数据分析、软件生命周期中的数据采集与分析、大规模群体协同的软件开发方法与平台环境、面向软件系统创新与产品线演化的模型、理论与工具、面向特定领域的大数据应用需求分析、用户行为数据收集与系统可用性分析、软件数据分析的代价与价值评估.

专题公开征文,共征得投稿 22 篇.特约编辑邀请了国内外在该领域有影响力的学者参与审稿工作,每篇投稿至少邀请 2 位专家进行初审.大部分稿件经过初审和复审两轮评审,部分稿件经过了两轮复审.通过初审的稿件还在 NASAC 2016(第 15 届全国软件与应用学术会议)大会上进行了现场报告,作者现场回答问题,并听取了听众的修改建议.最终有 13 篇论文入选本专题.入选论文覆盖开源软件生态分析、软件分类检测与推荐、软件故障与缺陷预测、面向领域的大数据应用开发方法.

在互联网环境下,以开源软件生态为代表的、以群体化协同模式开发的、深度结合用数据分析技术的软件系统成为软件研究新主题.

杨波等人《GitHub 开源软件开发过程中影响因素的相关性分析》的工作通过分析 GitHub 开源软件的开发过程,提出了问题解决速度、问题增加速度等影响因素,并对这些影响因素间的相关性进行了分析.

张宇霞等人的《OpenStack 开源社区中商业组织的参与模式》采用雪球采样方法对 OpenStack 相关的文本数据进行收集、过滤和归纳,总结出不同商业组织参与 OpenStack 的模式,为商业组织参与开源项目提供经验参考与决策支持.

杨程等人的《基于多维特征的开源项目个性化推荐方法》从开源项目自身流行度、关联项目技术相关度以及大众贡献者之间的社交关联度这 3 个维度度量开发者和开源项目之间的关联关系,并利用线性组合等方法构建推荐模型,从而为开发者提供个性化的项目推荐服务.

收稿时间: 2016-11-27; jos 在线出版时间: 2016-11-29

CNKI 网络优先出版: 2016-11-29 13:35:11, http://www.cnki.net/kcms/detail/11.2560.TP.20161129.1335.013.html

随着软件工程规模和项目数量的快速增长,软件制品的检测与分类问题成为软件工程领域的热点问题.

王浩宇等人的《大规模移动应用第三方库自动检测和分类方法》提出第三方库自动检测和分类方法,采用多级聚类技术和机器学习方法对第三方库进行识别和分类,并对 130 000 个 Android 应用进行实验分析.

徐培兴等人的《一种面向软件配置管理制品的层次分类方法》提出一种面向 CMT 制品的基于在线非结构化描述文档分析的层次分类方法.该方法利用标签共现性关系建立层次类别体系,基于描述属性特征,实现对 CMT 制品的层次分类器;并使用混合的样本划分方式针对数据倾斜问题进行了改进.对超过 11 000 例训练数据和 1 000 例测试数据进行了实验.

软件代码的搜索、理解和提交是目前学术界和工业界共同关注的热点研究主题.

黎宣等人的《基于增强描述的代码搜索方法》提出了一种基于增强描述的代码搜索方法 DERECS.该方法基于开源项目、问答系统等,构建一个代码-描述语料库,并分析代码及自然语言描述,提取方法调用和代码结构相关特征值,然后基于代码片段中的方法调用及代码片段的结构特征对代码进行描述增强,以减小被搜索的代码与自然语言查询语句之间的差异,扩大搜索的范围.

黄袁等人的《基于关键类判定的代码提交理解辅助方法》通过对大量数据的分析发现,识别出提交中关键及依赖类,能够辅助代码提交的理解,据此提出一种基于机器学习的关键类识别方法,将判定提交中的关键类建模为二分类问题.

软件质量、可靠性和安全性是软件工程研究的重要问题.

王子勇等人的《一种基于执行轨迹监测的微服务故障诊断方法》提出一种基于执行轨迹监测的微服务故障诊断方法.利用动态插桩监测服务组件的请求处理流,利用调用树对请求处理的执行轨迹进行刻画,利用树编辑距离来评估请求处理的异常程度,分析执行轨迹差异来定位引发故障的方法调用,并采用主成分分析抽取引起系统性能异常波动的关键方法调用.

何吉元等人的《一种半监督集成跨项目软件缺陷预测方法》提出了一种基于搜索的半监督集成跨项目软件缺陷预测方法 S^3EL.该方法首先通过调整训练集中各类数据的分布比例,构建出多个朴素贝叶斯基分类器,随后利用具有全局搜索能力的遗传算法,基于少量已标记目标实例对上述基分类器进行集成,并构建出最终的缺陷预测模型.在 Promise 数据集及 AEEEM 数据集上和多个经典的跨项目缺陷预测方法进行了对比.

蔡维德等人的《基于区块链的应用系统开发方法研究》从技术层面及应用层面分析区块链的特征,对区块链的设计需求、区块链的一致性和可扩展性进行了深入分析.给出了区块链的应用系统开发思路,以及账户区块链和交易区块链的双链设计模型,提出了链上代码并行执行模型应用原则,并对区块链技术应用进行总结和展望.

俞一峻等人的《小模型大数据:一种分析软件行为的代数方法》通过问题框架方法分析软件需求领域知识的结构化建模,诠释需求的可满足性.需求可满足性的定性分析及争辨能够支持早期设计决策,选择合理的软件体系结构和设计.该文从软件抽象目标行为的角度深化问题框架的分析思路,针对特定行为建立相应的概率描述模型,提出一种基于代数的分析方法,以弥补纯粹数据驱动思路的大数据分析盲点.通过对安全和隐私性需求实例的探讨,对大数据软件系统需求研究给出启示.

海量数据的搜集、存储、分析、处理和利用成为软件系统的热点应用.

朱美玲等人的《基于车牌识别流数据的车辆伴随模式发现方法》针对新兴的智能交通应用中车牌识别流式大数据,将 Platoon 伴随模式发现问题映射为数据流上的带有时空约束的频繁序列挖掘问题,提出 PlatoonFinder 算法在车牌识别数据流上即时挖掘 Platoon 伴随模式.算法在序列挖掘的过程中有效处理频繁序列元素之间复杂的时空约束关系;并融入了伪投影等性能优化技术.

王建民的《领域大数据应用开发与运行平台技术研究》为降低大数据技术在各行各业应用普及的门槛,为面向领域的大数据应用系统的快捷开发和高效运行提供方法、工具和平台支撑.该文指出,由于大数据固有的复杂性、动态性、多样性及其价值独创性,目前尚未形成系统化的大数据软件开发方法,难以满足不同领域对大数据全生命周期处理的多样化需求.大数据时代的软件工程面临的挑战,体现在互为依赖的两方面:面向大

数据全生命周期的集成设计开发环境和基于软件生命周期的系统运行分析工具.结合特定领域的实际需求,该文探讨研究面向领域的大数据应用系统开发与运行一体化平台技术,覆盖大数据生命周期(获取、清洗、集成、分析、呈现)以及软件生命周期(设计、开发、运行、优化),形成自管理、自适应、自优化的平台化解决方案.

本专题主要面向软件工程研究与应用领域的相关人员,反映了我国学者在该领域的最新研究进展.在此,我们要特别感谢《软件学报》编委会对专题工作的指导和帮助,感谢编辑部各位老师从征稿启示发布、审稿专家邀请至评审意见汇总、论文定稿、修改及出版所付出的辛勤工作和汗水,感谢专题国内外评审专家及时、耐心、细致的评审工作.此外,我们还要感谢向本专题踊跃投稿的作者对《软件学报》的信任.

最后,感谢专题的读者们,希望本专题能够对相关领域的研究工作有所促进.

刘璘(1973－),女,清华大学软件学院副教授,研究生导师.CCF 专业会员,CCF 软件工程专委会委员.主要研究领域为需求工程,数据与知识工程,领域工程.

周明辉(1974－),女,北京大学信息科学技术学院副教授,研究生导师.CCF 专业会员,CCF 软件工程专委会委员.主要研究领域为软件度量,软件数据挖掘,软件数字考古学.

尹刚(1975－),男,国防科学技术大学计算机学院副教授,研究生导师.CCF 专业会员,CCF 软件工程专委会委员.主要研究领域为软件工程,机器学习,软件教学与实训.