

专题：数据驱动的软件智能化开发

Data-Driven Intelligent Software Development

客座编辑



谢冰(1970-)，男，博士，北京大学教授、信息科学技术学院常务副院长、软件研究所所长，国家杰出青年基金获得者，中国软件行业协会理事，中国计算机学会高级会员，*Chinese Journal of Electronics*编委，入选教育部新世纪优秀人才支持计划、北京市科技新星计划，获得“中创软件人才奖”。主要研究方向为软件工程、计算机理论科学和分布式系统等。在软件工程技术研究和国产软件工具研究与产业化方面作出了重要的贡献，主要成果在软件行业、信息服务业、集成电路设计领域得到了广泛应用。在软件工程与程序语言领域顶级会议(FSE、POPL、ASE等)和国际期刊上发表论文80余篇。曾获得2006年国家科技进步奖二等奖(排名第三)、2009年北京市科学技术奖技术发明类二等奖(排名第一)、2012年教育部高等学校科学技术进步奖一等奖(排名第一)、2015年国家技术发明奖二等奖(排名第二)。



彭鑫(1979-)，男，博士，复旦大学教授、计算机科学技术学院副院长、软件学院副院长。中国计算机学会软件工程专业委员会副主任，*Journal of Software: Evolution and Process*联合主编，*ACM Transactions on Software Engineering and Methodology*编委，《软件学报》编委，*Empirical Software Engineering*编委，IEEE软件维护与进化国际会议(ICSME)执行委员(2017—2020年)。2016年获得NASAC青年软件创新奖。主要研究方向为软件开发大数据分析、智能化软件开发、云原生与智能化运维、泛在计算软件系统等。在ICSE、FSE、ASE、CSCW、ICSME、SANER、RE等软件工程及相关领域高水平国际会议及IEEE/ACM Transactions等国际期刊发表论文80余篇。研究工作获得ICSM 2011最佳论文奖、ACM SIGSOFT杰出论文奖(ASE 2018)、IEEE TCSE杰出论文奖(ICSME 2018/2019/2020)、*IEEE Transactions on Software Engineering*年度最佳论文奖(2018年)。担任ICGSE 2014及ICSME 2017组委会主席、Internetware 2017及ICSR 2019程序委员会主席，以及ICSE、FSE、ASE、ICSME、SANER、RE等会议程序委员会委员。带领复旦大学CodeWisdom研究团队开展软件开发大数据分析平台以及智能化软件开发平台的研究与实践，研究成果在多家大型企业进行了实践应用。

导读

以开源软件为代表的互联网软件开发呈现了边界开放、群体分散、交付频繁、知识复杂等特征。与此同时,企业软件开发也逐渐转向以开发运维一体化(DevOps)为特征的云化开发平台。这种网络化开发方式产生了包含源代码、缺陷报告、版本历史、测试用例、部署信息及运行日志在内的全生命周期数据,其中蕴涵着大量的软件开发知识。智能化软件开发一直是软件工程追求的核心目标之一。随着软件开发大数据的不断积累,数据驱动的智能软件开发技术研究已经成为学术界和工业界共同关注的热点问题。然而,软件开发大数据的汇聚组织、知识提炼、软件工具智能化和智能开发服务环境等仍然是亟待解决的问题。本专题汇集了国内智能化软件开发领域几支活跃的研究团队的7篇文章。这些团队在国家重点研发计划项目“基于大数据的软件智能开发方法和环境”的支持下,在智能化软件构造、测试验证、群体协作、运维演化4个方面针对软件智能化开发服务进行了深入的研究。本专题反映了这些团队在这几个方面的研究成果以及对相关技术发展方向的观点。

谢冰等人撰写的《基于大数据的软件智能化开发方法与环境》阐述了项目团队围绕软件工程大数据的汇聚组织、知识表示提炼、软件工具智能化和智能开发服务环境等关键技术开展的研究工作,以及构建的软件智能化开发环境和服务平台。

邹艳珍等人撰写的《基于大数据的软件项目知识图谱构造及问答方法》阐述了以代码结构为核心的软件知识图谱模型以及“知识抽取-知识融合”两层软件知识图谱构造框架,提出了软件项目知识图谱自动构造方法以及基于知识图谱的软件项目智能问答方法。

彭鑫等人撰写的《基于上下文的智能化

代码复用推荐》围绕基于上下文的智能化代码复用推荐这一问题,阐述了基于模板挖掘的代码复用推荐和基于深度学习的代码复用推荐两方面的研究工作,并对基于上下文的智能化代码复用推荐的未来发展方向进行了展望。

席圣渠等人撰写的《基于大数据的开源项目缺陷报告智能预检技术》提出了一种智能化缺陷报告预检技术框架,归纳了缺陷报告预检中存在的3个关键任务(缺陷优先级分类、缺陷分派、缺陷再分派),同时结合开源项目的特点介绍了相关技术。

陈伟等人撰写的《基于领域知识的Docker镜像自动构建方法》提出了一种基于领域知识的Docker镜像自动构建方法,并通过实验验证了该方法具有利用领域知识推断系统依赖关系和软件包安装方式、生成不同软件Dockerfile的能力。

张建等人撰写的《数据驱动的软件开发者智能协作技术》阐述了海量软件开发者数据的收集和分析方法,定义了软件开发者能力特征模型及协作关系模型,介绍了开发者知识图谱构建方法、基于智能推荐的协作开发方法以及所构建的智能协作开发环境。

张洋等人撰写的《面向智能化软件开发的开源生态大数据》通过对软件工程开源生态大数据体系的研究提出了一种自生长的采集处理框架与汇聚共享环境,阐述了基于软件工程开源生态大数据的智能化软件开发相关研究以及相关的分析挖掘方法典型应用。

由于篇幅有限,本专题无法覆盖数据驱动的软件智能化开发的各个方面,希望通过分析当前技术现状、介绍已有研究进展、展望未来发展方向,阐述该领域面临的机遇和挑战,引起学术界和产业界的关注,促进相关研究的发展以及研究成果在实践中的应用。