Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

社论：

软件数据智能分析：研究与应用

张涛[1]，孙小兵[2]，郑子彬[3]，李戈[4]

[1]澳门科技大学计算机科学与工程学院，中国澳门特别行政区

[2]扬州大学信息工程学院，中国扬州市，225127

[3]中山大学软件工程学院，中国珠海市，519082

[4]北京大学计算机学院，中国北京市，100871

E-mail: tazhang@must.edu.mo; xbsun@yzu.edu.cn; zhzibin@mail.sysu.edu.cn; lige@pku.edu.cn

本文编译自 Zhang T, Sun XB, Zheng ZB, et al., 2022. Intelligent analysis for software data: research and applications. *Front Inform Technol Electron Eng*, 23(5):661-663. https://doi.org/10.1631/FITEE.2230000

软件一直是世界经济增长的主要动力之一，人类生活依赖于可靠的软件。软件生产过程（即软件设计、开发、测试和维护）是确保软件质量的最重要因素之一。在生产过程中，会产生大量软件数据（例如，源代码、错误报告、日志和用户评论）。随着软件复杂性增加，如何利用软件数据来提高软件生产性能和效率，成为软件开发者和研究人员的一个挑战。为应对这一挑战，研究人员利用信息检索、数据挖掘和机器学习技术，实现了一系列自动化工具，以提高一些重要软件工程任务的效率，如代码搜索、代码摘要生成、严重性/优先级预测、缺陷定位和程序修复。然而，这些传统方法不能深入捕捉上下文信息的语义关系，而且通常忽略了源代码的结构信息。因此，这些自动化软件工程任务的性能仍有提高余地。

“智能”一词意味着我们可以使用新一代人工智能技术（如深度学习）来设计一系列“智能”自动化工具，以提高软件工程任务的有效性和效率，从而大大减少开发人员的工作量。这里，我们展示使用“智能”分析技术来解决两个经典的自动化软件工程任务。

1. 智能软件开发

代码搜索和代码摘要生成可以帮助开发人员开发高质量软件，并提高工作效率。代码搜索是软件开发中的日常操作，它可以帮助开发者找到合适的代码片段来完成软件项目。开发人员通常将这些代码片段的描述作为查询输入，以此来找到相应代码。然而，设计一个实用的代码搜索工具极具挑战性。以前基于信息检索的方法忽略了自然语言表达的高层描述和低层源代码之间的语义关系，影响了代码搜索性能。与基于信息检索的方法不同，深度学习技术可以自动学习特征表征，并在输入和输出之间建立映射关系。因此，代码搜索性能得到改善。代码摘要生成是自动生成源代码的自然语言描述的任务，它可以帮助开发人员理解和维护软件。在传统的自动化代码摘要生成工作中，研究人员倾向于使用总结出的模板来提取源代码关键词，这忽略了源代码的语法信息。目前，神经网络技术蓬勃发展，卷积神经网络（convolutional neural networks, CNN）、循环神经网络（recurrent neural networks, RNN）、Transformers 和其他深度学习网络被应用于代码摘要自动生成任务。

2. 智能软件维护

严重性/优先级预测可以自动推荐合适的标签，帮助开发人员减少标注严重性和优先级的工作量，这也是错误报告（bug report）的重要特征。严重性显示了报告中错误的重要性或严重程度，

‡ 通讯作者

而优先级则表明哪些错误应该首先被修复。预测任务可以帮助开发人员快速将重要的错误分配给合适的开发人员修复，从而提高软件维护效率。传统方法通常采用机器学习技术，如支持向量机（support vector machine, SVM）和朴素贝叶斯（Naive Bayes, NB），来预测严重性/优先级。然而，这些方法无法克服数据不平衡问题，所以预测准确性并不完美。一些深度学习技术，如 CNN 和图卷积网络（graph convolutional networks, GCN），可以有效解决该问题，通过捕获错误报告的上下文语义信息，提高预测性能。

在此背景下，中国工程院院刊《信息与电子工程前沿（英文）》组织了本期关于软件数据智能分析的专题。专题涵盖了软件架构恢复、应用审查分析、集成测试、软件项目管理、缺陷预测和方法重命名以及相关应用。经严格评审，选入 6 篇研究论文。

软件架构在软件生命周期中发挥着重要作用，特别是在软件演化中。维护最新架构文件是很困难的，因为它应该包含所有软件利益相关者信息。因此，软件架构恢复任务的目的是从软件系统的低级表示（如源代码）中识别和提取体系结构信息。然而，这一任务在学术界和工业界都代价高昂。为解决该问题，李必信等提出基于现有架构和相关代码变化的增量式软件架构恢复技术，（即 ISAR）。他们建立了代码级变化和架构级更新之间的映射，帮助研究人员提高软件恢复技术的性能。基于 10 个开源项目的评估结果表明，ISAR 性能优于传统方法。

在用户评论中准确识别新兴主题（如软件漏洞）可帮助开发者更有效地更新应用。然而，由于用户评论文本长度较短、提供的信息有限，新兴主题识别的准确率较低。为解决该问题，王勇等提出一种改进的新兴主题识别方法。他们采用自然语言处理技术减少噪音数据，并使用自适应在线双词主题模型（AOBTM）识别应用程序评论中的新兴主题。实验结果表明，所提方法能有效识别新兴主题。

集成测试是软件测试的重要组成部分。现有生成测试用例的方案主要关注减少集成测试序列生成的成本，未考虑赋予可靠性风险较大的节点较高测试优先级。于海等提出多层动态执行网络（multilayer dynamic execution network, MDEN）模型，利用概率风险评估方法为软件中每一个类量化测试优先级。此外，提出一种优化策略，在生成测试用例的过程中保证两条原则：一是为高风险的类赋予较高权重，二是最小化测试桩复杂度。与现有算法的实验对比分析证明所提方法优于基线方法。

跨项目软件缺陷预测解决了传统缺陷预测中训练数据不足的问题，但仍然存在两个挑战：（1）模型训练过程中，过多无关和冗余特征影响了训练效率，降低了模型预测精度；（2）由于开发环境不同，度量值的分布因项目而异，导致模型在跨项目预测时精度较低。为解决这两个难题，蔡赛华等提出一种基于特征选择和迁移学习的软件缺陷预测方法。实验结果表明，所提方法表现出较好性能。

程序中的方法必须被准确命名，以方便源代码分析和理解。在软件开发过程中，方法体可能变得与方法名称不一致，导致方法名称不准确或出现错误。以往研究集中在当方法体被修改时推荐准确的方法名称；然而，这些研究存在两个问题：（1）缺乏对方法名称结构的分析；（2）缺乏对编程环境上下文信息的有效捕获。为解决这些问题，张静宣等提出一种新的方法重命名算法，利用结构和词法分析推荐高质量的方法名称。他们开展了一系列实验来验证所提方法的有效性，结果表明该方法可显著改善最先进的方法。

面向任务的虚拟助手是为用户提供自然语言界面以完成特定领域任务的软件系统。然而，由于自然语言理解问题的复杂性和困难性，管理一个面向任务的虚拟助手软件项目具有挑战性。李姝玥等分享了解决这些问题的做法，介绍了在管理面向任务的虚拟助手软件项目方面的经验教训。他们还开发了一个新的需求管理工具以提高面向任务的虚拟助手软件项目的管理效率。

总之，上述 6 项研究涵盖了许多自动化任务，通过分析软件数据来提高软件开发的有效性和效率。此外，提供了一系列解决方案来克服之前研究中的挑战。我们希望这些课题对软件数据智能分析及相关领域研究人员有所帮助。

最后，我们要特别感谢作者和审稿人对本专题的支持和宝贵贡献，感谢编辑部工作人员和主编潘云鹤院士、卢锡城院士。

张涛，于东北大学获自动化学士学位和软件工程硕士学位，韩国首尔市立大学获计算机科学博士学位，之后在香港理工大学进行博士后研究，现供职于澳门科技大学计算机科学与工程学院。是 IEEE 和 ACM 高级会员。发表软件工程和安全相关论文 60 多篇，包括 *IEEE Trans Softw Eng*、*IEEE Trans Inform Forens Sec*、*IEEE Trans Depend Sec Comput*、*IEEE Softw* 等期刊论文以及 ICSE 等会议论文。研究兴趣包括软件仓库挖掘和移动软件安全。

孙小兵，2007 年于江苏科技大学获计算机科学与技术学士学位，2012 年于东南大学获博士学位，现为扬州大学信息工程学院教授。在国际权威期刊（*EMSE*、*STVR*、*IST*、*JSS*、*SCIS*、*FCS* 等）和会议（ICSE、ASE、ICSME、SANER、ICPC 等）发表论文 80 余篇。研究兴趣包括软件仓库挖掘和智能分析、软件安全等。

郑子彬，2011 年于香港中文大学获博士学位，现担任中山大学软件工程学院教授。发表 150 多篇国际期刊和会议论文，包括 3 篇 ESI 高被引论文。曾获 ACM 杰出论文奖。研究兴趣包括区块链、智能合约、服务计算和软件可靠性。

李戈，北京大学计算机学院副教授。1999 年获山东理工大学学士学位，2006 年获北京大学博士学位，2013~2014 年，在美国斯坦福大学人工智能实验室担任客座副教授。是 CCF 软件工程学会副秘书长和软件程序生成研究小组创始人，该小组包括中国 100 多名高级研究人员。目前研究主要涉及机器学习概率方法的应用，包括程序语言处理、程序代码生成和自然语言处理。