基于近似子树匹配的快速代码推荐方法

邵宜超[1,2,3]，黄志球[1,2,3]，李伟漳[1,2,3]，喻垚慎[1,2,3]
[1]南京航空航天大学计算机科学与技术学院，中国南京市，211100
[2]工业和信息化部安全关键软件重点实验室，中国南京市，211100
[3]软件新技术与产业化协同创新中心，中国南京市，210016

摘要：软件开发人员通常需编写与已有代码具有类似功能的代码，而帮助开发人员重用这些代码片段的代码推荐工具可显著提高软件开发效率。近年来许多研究者开始关注这一领域，并提出多种代码推荐方法。一些研究者使用序列匹配算法得到相关代码，这些方法往往效率较低，且只能利用代码中的文本信息。另一些研究者从代码中提取特征并形成特征向量，从而计算代码间相似性并得到推荐结果。然而特征向量相似往往不代表原始代码相似，在将抽象语法树转换为向量的过程中存在结构信息丢失问题。对此，我们提出一种基于近似子树匹配的代码推荐方法。与现有基于特征向量匹配的方法不同，该方法在匹配过程中保留了查询代码的树型结构，从而找到与当前查询在结构上最为相似的代码片段。此外，通过哈希思想将子树匹配问题转化为树与列表间的匹配，使得抽象语法树信息可以用于对时间要求较高的代码推荐任务。为评估方法的有效性，构建了多个涵盖不同语言和粒度的代码数据集。实验结果表明，该方法在所有数据集上的召回率均优于两种对比方法——SENSORY和Aroma，且可以应用于大型数据集。

关键词：代码复用；代码推荐；树相似度；结构信息
https://doi.org/10.1631/FITEE.2100379