ОТЗЫВ

официального оппонента, доктора физико-математических наук Галатенко Владимира Антоновича на диссертационную работу Саргсяна Севака Сениковича «Методы поиска клонов кода и семантических ошибок на основе семантического анализа программы», представленную к защите на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11– «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Актуальность. В процессе разработки программного обеспечения (ПО) обычной практикой является повторное использование (клонирование) фрагментов исходного кода. Разработчики могут копировать фрагменты кода, написанные ими самими, или заимствовать их из других источников, например, из ПО с открытыми исходными текстами. Клонирование экономит усилия программистов, но оно также может быть причиной увеличения размера кода и источником семантических ошибок, вызванных некорректной адаптацией скопированных фрагментов кода. В связи с этим в настоящее время ведутся активные исследования в области разработки инструментов для выявления клонов кода и связанных с ними семантических ошибок. Внедрение подобных инструментов в цикл разработки ПО может способствовать снижению затрат на создание и сопровождение ПО, а также повышению его надежности. Таким образом, тема диссертации С.С. Саргсяна, посвященной методам поиска клонов кода и семантических

ошибок в них, является актуальной.

Структура и содержание диссертации: диссертация состоит из введения, пяти глав, заключения и списка литературы, который включает 98 наименований.

Во введении обосновывается актуальность темы диссертации и формулируется цель исследования.

В первой главе приведен обзор работ по теме диссертации. Дается сравнительный анализ существующих методов поиска клонов кода и семантических ошибок.

Во второй главе рассматривается разработанный автором метод поиска клонов кода на основе семантического анализа программы, включающий четыре фазы: (1) разделение графа зависимостей программы (ГЗП) на подграфы; (2) отсеивание несхожих пар ГЗП; (3) поиск максимальных схожих подграфов в заданной паре ГЗП; (4) фильтрация ложных срабатываний. Представлены новые алгоритмы для решения следующих задач: разделение ГЗП на подграфы; отсеивание несхожих пар подграфов ГЗП; поиск клонов кода на основе предложенной автором метрики кода, изоморфизма деревьев и слайсинга. Эффективность алгоритмов подтверждается результатами тестирования на ряде больших проектов с открытыми исходными текстами (Linux 2.6, Firefox Mozilla, LLVM/Clang, OpenSSL).

В третьей главе описана архитектура инструмента поиска клонов кода, включающая три основных компонента. Первый из них выполняет построение ГЗП из промежуточного представления LLVM на этапе компиляции программы и их сохранение. ГЗП могут генерироваться с разной степенью детализации в зависимости от решаемой задачи. Второй компонент выполняет поиск клонов кода. Третий компонент предназначен для отладки и

тестирования алгоритмов поиска клонов. Он автоматически генерирует набор клонов кода для заданного проекта и запускает алгоритм поиска. Высокая точность разработанного автором инструмента подтверждается результатами тестирования на проектах Linux 2.6, Firefox Mozilla, LLVM/Clang, OpenSSL. В главе 3 также представлены средства параллельного запуска инструмента в многоядерных системах и средства для просмотра результатов поиска клонов.

В четвертой главе представлен метод поиска клонов кода в скриптах на языке JavaScript. Рассматривается реализованное автором дополнение в динамическом компиляторе V8 языка JavaScript, выполняющее генерацию ГЗП из промежуточного представления Hydrogen. Поиск клонов на основе ГЗП осуществляется при помощи инструментов, описанных в главе 3. Анализ продуктивности разработанных средств проводился на известных тестовых наборах SunSpider, Octane, Kraken. Число выявленных клонов составило 342, 10, 0, соответственно; из них более 90% являются истинными. На наборе SunSpider число найденных клонов в 10 раз превышает число клонов, выявленных системой CloneDR.

В пятой главе рассматривается задача поиска семантических ошибок, возникающих в результате неполной адаптации повторно использованных фрагментов кода. Предложен комбинированный метод, использующий одновременно лексический и семантический анализ программы. Вначале путем лексического анализа программы выявляются совпадающие фрагменты исходного кода и для них строятся ГЗП графы. Поиск семантических ошибок осуществляется путем анализа полученных ГЗП. Применение этих средств к ряду больших проектов с открытым исходным кодом (Android 4.3, Linux 2.6, Firefox Mozilla, LLVM/Clang, OpenSSL, Qemu) позволило выявить более ста ошибок, связанных с некорректным

переименованием переменных.

В заключении подводятся итоги проделанной работы и формулируются направления дальнейших исследований.

Основные результаты диссертации, обладающие научной новизной.

1. Масштабируемый четырехфазный метод поиска клонов кода на основе семантического анализа программы. Реализация метода включает разработанные автором новые алгоритмы для следующих задач: разделение ГЗП на подграфы; отсеивание несхожих пар подграфов ГЗП; поиск максимальных схожих пар подграфов на основе слайсинга, метрики и изоморфизма деревьев; фильтрация ложных клонов.
2. Метод поиска семантических ошибок на основе комбинированного лексического и семантического анализа.
3. Расширяемая архитектура инструмента поиска клонов кода для языков программирования C, C++ и JavaScript.
4. Подсистема анализа и тестирования алгоритмов поиска клонов, предназначенная для проверки точности реализованных алгоритмов с целью их дальнейшего совершенствования.
5. Реализованные автором инструментальные средства: масштабируемый инструмент поиска клонов кода на базе компиляторной инфраструктуры LLVM; генератор ГЗП в JIT-компиляторе V8 для осуществления поиска клонов кода в скриптах на языке JavaScript; инструмент поиска семантических ошибок.

Практическая значимость: разработанные инструментальные средства используются в научно-исследовательских проектах Института системного программирования Российской академии наук (ИСП РАН). Выявление клонов кода и семантических ошибок в процессе разработки ПО

способствует повышению качества ПО, а также снижению затрат на его сопровождение.

Апробация. По теме диссертации опубликовано 7 статей, 4 из которых в изданиях, входящих в перечень ВАК РФ. Результаты диссертации представлены на следующих международных конференциях: Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2014»; 57-я научная конференция МФТИ, 2014 г.; FOSDEM-2015; CSIT-2015; Открытая конференция по компиляторным технологиям, 2015 г.

Замечания. По диссертационной работе можно сделать следующие замечания:

1. В главе 1 не приведено формальное определение клона кода. Дается лишь «интуитивное» определение путем выделения трех типов клонов.
2. В главе 2 не приведен анализ алгоритмической (не)разрешимости предлагаемых методов поиска клонов кода.
3. В главе 2 не рассмотрены особенности выявления клонов, возникающих при использовании макросов, встраивании функций (англ. inlining) и шаблонов (англ. templates).
4. В диссертации не описывается методика применения результатов.
5. В оформлении работы имеются погрешности: отсутствует отдельный пункт «Научная новизна», допущен ряд опечаток (например, на стр. 8, 12 диссертации).

Перечисленные замечания не носят принципиального характера и не влияют на общую положительную оценку работы.

Заключение: Диссертационная работа соответствует всем требованиям ВАК РФ, предъявляемым к диссертациям на соискание ученой степени кандидата физико-математических наук, а Саргсян Севак Сеникович

заслуживает присуждения ему ученой степени по специальности 05.13.11 – «математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Заведующий сектором ФГУ ФНЦ НИИСИ РАН,
доктор физико-математических наук В.А. Галатенко

Подпись ру
Начальник с

«25» февраля 2016 г.