УДК 004.89, 004.4'4,
004.45

ОБНАРУЖЕНИЕ ПЛАГИАТА В ИСХОДНОМ КОДЕ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

Аль Али Моаз (Университет ИТМО)
Научный руководитель – к.ф.-м.н., доцент Фильченков А.А.
(Университет ИТМО)

Предлагается решение на основе нейронных сетей для обнаружения плагиата в программном коде, которое позволит избежать попарного сравнения между всеми возможными парами исходного кода с целью улучшения производительности автоматических средств проверки на плагиат с точки зрения времени выполнения и точности.

Введение. Плагиат — распространенное явление во многих областях, особенно в области письменной литературы. Оно определяется как несанкционированное использование чужой работы как собственной работы без указания первоначального источника информации. Плагиат — большая проблема, особенно в академической сфере, которая может привести к недостатку знаний или навыков, а также к несправедливой оценке по отношению выпускниками академических учреждений. В последние несколько лет из-за быстрого развития в области информатики и легкого доступа к большому количеству фрагментов общедоступного исходного кода в Интернете на некоторых сайтах вопросов и ответов (question-and-answer sites) таких как Stack Overflow, Quora и других появился новый тип плагиата, который является плагиатом исходного кода. Очень сложно сравнивать все возможные пары файлов исходного кода разных программистов, это отнимает много времени из-за большого количества возможных программистов в академических кругах, на конкурсах программирования и т. д. Таким образом, возникает необходимость в разработке точных и высокоэффективных автоматических методов обнаружения плагиата в исходном коде.

Большинство существующих методов построены на попарных операциях между исходными кодами с различными подходами, как:

1. Преобразование исходного кода в последовательности токенов и попарное сравнение кодов с использованием алгоритмов сопоставления строк.
2. Использование абстрактного синтаксического дерева (AST) для получения признаков для представления исходного кода.
3. Использование идеи n-грамм (n-grams) токенов или узлов AST для представления кода.
4. Использование семантического анализа для извлечения признаков, определяя отношения между частями бинарного кода и базовыми блоками кода.
5. Решения на основе информационного поиска, построенные на извлечении признаков, представляющие код, их индексировании, и преобразовании подозрительного кода в виде запроса системы на наличие аналогичных кодов.

Основная часть. Предлагаемый метод основан на идее представления исходного кода в виде векторов вещественных чисел, которые представляют собой набор выбранных признаков разных типов (лексических, структурных и синтаксических признаков) из исходного кода. Эти признаки извлекаются из самого исходного кода, из бинарного кода после компиляции и из абстрактного синтаксического дерева кода. После извлечения признаков применяется выделение признаков, с целью сохранения только признаков с наибольшим приростом информации. Чтобы избежать попарного сравнения всех кодов, только исходные коды, которые близки друг к другу, помещаются в одну группу для подачи в виде пар в классификатор на основе нейронной сети. Коды группируются по расстоянию в пространстве признаков с использованием структуры B-дерева.

Выводы. Результаты данной работы могут быть использованы в академических классах программирования для обнаружения плагиата между решениями учащихся или в соревнованиях по программированию для предотвращения мошенничества и гарантии справедливых результатов. Кроме того, обнаружение плагиата между вредоносным кодом и кодом конкретного программиста может привести к выявлению того факта, что этот программист является инициатором вредоносной программы.

Аль Али Моаз (автор) Подпись

Фильченков А.А. (научный руководитель) Подпись