МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ» (ТУСУР)

Кафедра комплексной информационной безопасности электронно-вычислительных систем (КИБЭВС)

ОПРЕДЕЛЕНИЕ АВТОРСТВА ИСХОДНОГО КОДА

ОТЧЕТ

по преддипломной практике

Студентка гр. 722

_____ Мейта М.В.

«____»_____2017г.

Руководитель практики

к.т.н., доцент каф. БИС

_____ Романов А.С.

«____»_____2017г.

Томск 2017

РЕФЕРАТ

Отчет содержит 15 страниц, 2 рисунков, 0 таблиц, 7 источников, 1 приложение.

СТИЛОМЕТРИЯ, ИСХОДНЫЙ КОД, ДЕАНОНИМИЗАЦИЯ АВТОРА, C++, КЛАССИФИКАЦИЯ, PYTHON, SKLEARN, RANDOM FOREST CLASSIFIER, LATEX.

Цель работы — разработка алгоритма для определения авторства программного обеспечения, основанного на стилометрическом анализе исходного кода программ на языках высокого уровня.

В рамках научно-исследовательской работы на текущий семестр были поставленны следующие задачи:

– обзор существующих исследований, разработок, методов стилометрического анализа тек- ста, в том числе исходного кода программ;

– разработка алгоритм анализа исходного кода программ с применением стилометрии для определения авторства программного обеспечения;

– создание программной реализации разработанного алгоритма;

– исследование эффективности разработанного алгоритма анализа исходных кодов.

Объект исследования: деанонимизация автора программного обеспечения.

Предмет исследования: стилометрия исходного кода программ на языках высокого уровня.

В результате работы было выполнено слудующее:

– произведен аналитический обзор существующих методов анализа исходного кода программ с целью деанонимизации автора;

– выбран набор признаков для классификации авторов программного кода на языке C++;

– в качестве алгоритма классификации выбран Random Forest Classifier;

– разработан алгоритм определения авторства исходного кода программ на языке C++;

– произведена программная реализация разработанного алгоритма на языке программирования высокого уровня Python;

– подготовлен тестовый набор данных;

– выбраны критерии оценки эффективности разработанного алгоритма;

– произведены вычислительные эксперименты на данном наборе данных;

– сделаны некоторые выводы на основе полученных результатов.

Отчет о НИР выполнен согласно ОС ТУСУР 01-2013 [1] при помощи системы компьютерной вёрстки LaTeX.

Содержание

Введение 5
1 Кафедра КИБЭВС 5
2 Обзор информационных источников 7
3 Моделирование 10
Заключение 11
Список использованных источников 13
Приложение А Компакт-диск 15

Ведение

С распространением применения компьютерных систем и сетей возросло и количество преступлений в информационной сфере. Существует множество разновидностей кибератак — различные компьютерные вирусы, трояны, несанкционированное копирование данных с кредитных карт, DDoS-атаки и многое другое. Возможность деанонимизации авторов вредоносного программного обеспечения может внести существенный вклад в развитие компьютерной криминалистики.

Целью данной работы является исследование методов стилометрии — статистического анализа текста для выявления его стилистических особенностей, а также методов машинного обучения для решения задачи деанонимизации разработчика по исходному коду программного обеспечения.

Определение авторства исходного кода представляет собой актуальную задачу в сфере информационной безопасности, лицензирования в области разработки программного обеспечения, а также может оказать существенную помощь во время судебных разбирательств, при решении вопросов об интеллектуальной собственности и плагиате.

1 Кафедра КИБЭВС

Местом прохождения практики была выбрана кафедра ТУСУРа – КИБЭВС (Кафедра комплексной информационной безопасности электронно- вычислительных систем).

Кафедра организована в ТУСУР в 1971 году как кафедра «Конструирования и производства электронно-вычислительной аппаратуры» (КиПЭВА) вскоре переименованной в кафедру «Конструирования электронно-вычислительной аппаратуры» (КЭВА).

21 сентября 1999 г. в связи с открытием новой актуальной специальности 090105 – «Комплексное обеспечение информационной безопасности автоматизированных систем» кафедра КЭВА была переименована в кафедру «Комплексной информационной безопасности электронно-вычислительных систем» (КИБЭВС). За-

ведующим кафедрой КИБЭВС на сегодняшний день является ректор ТУСУРа, Александр Александрович Шелупанов, лауреат премии Правительства Российской Федерации, действительный член Международной Академииnнаук высшей школы РФ, действительный член Международной Академии информации, Почетный работник высшего профессионального образования РФ, заместитель Председателя Сибирского регионального отделения учебно-методического объединения вузов России по образованию в области информационной безопасности, профессор, доктор технических наук.

С 2008 г. кафедра КИБЭВС входит в состав Института «Системной интеграции и безопасности».

На базе кафедры КИБЭВС ТУСУР в 2002 году организовано «Сибирское региональное отделение учебно-методического объединения Вузов России по образованию в области информационной безопасности [1].

Официальный сайт КИБЭВС [Электронный ресурс]. – Режим доступа: http://kibevs.tusur.ru/pages/kafedra/index (дата обращения:

2 Обзор информационных источников

На первом этапе научно-исследовательской работы необходимо было провести аналитический обзор информационных источников, рассмотреть существующие методы определения авторства исходного кода и различные подходы к решению такого рода задачи.

В работе [2] представлен набор инструментов и техник, используемых для решения задач анализа авторства исходного кода, а также обзор некоторых наработок в данной предметной области. Кроме того, авторы приводят собственную классификацию проблем и подходов к их решению в рамках задачи деанонимизации авторов программного обеспечения.

Среди проблем (задач) анализа авторства исходного кода выделены:

– идентификация автора — направлена на определение, принадлежит ли определенный фрагмент кода конкретному автору;

– характеристика автора — базируется на анализе стиля программирования;

– определение плагиата — нахождение схожестей среди множества фрагментов файлов исходного кода;

– определение намерений автора — был ли код изначально вредоносным или стал таковым в следствие программной ошибки;

– дискриминация авторов — определение, был ли код написан одним автором или несколькими.

Подходы к решению вышеперечисленных проблем (задач):

– анализ «вручную» — данный подход включает в себя исследование и анализ фрагмента исходного кода экспертом;

– вычисление схожести — базируется на измерении и сравнении различных метрик или токенов для набора файлов исходного кода;

– статистический анализ — в таком подходе используются статистические техники, такие как дискриминантный анализ и стилометрия, позволяющие определить различия между авторами;

– машинное обучение — используются методы рассуждения на основе пре-

цедентов и нейронные сети для классификации автора на базе некоторого набора метрик.

В работе [3] предложен способ определения авторства программного обеспечения. в основе которого лежит система, состоящая из 100 метрик, отражающих «почерк создателя» программного обеспечения. На основе метрик составлен «профиль почерка» пяти разных программистов по текстам трех разработанных ими программных систем и проверено соответствие этому профилю других программ, написанных в том числе и другими программистами. Однако авторы привели сомнительные результаты вычислений, не указали способ составления «профиля почерка» и полученную точность, с которой программная система определяла авторство.

В [4] исходный код транслировался в абстрактные синтаксические деревья, после чего разбивался на функции. Дерево каждой функции принималось за отдельный документ с известным автором. Выборка, состоящая из такого рода деревьев подавалась на вход SVM-классификатору, оперирующему данными типа «дерево». Классификатор обучался на файлах исходного кода двух авторов, в результате чего удалось достичь точности около 67-88%.

В статье [5] рассматривался способ атрибуции исходного кода с использованием метода N-грамм. Вопрос определения авторства программ в данной работе рассматривался с точки зрения определения плагиата. В качестве выборки использовался набор из 1640 файлов исходного кода, написанных 100 авторами. Производилось ранжирование документов по схожести, после чего производилась оценка результатов. При этом составителям удалось успешно определить плагиат в 67% случаев.

В [6] применялся алгоритм классификации Random Forest [7] и построение абстрактных синтаксических деревьев. Обучение и тестирование производилось для количества авторов от 250 до 1600. При этом удалось добиться высокой точности — 94-98%. Кроме того, авторы статьи выяснили в ходе работы, что сложнее определить авторов более простых примеров, нежели сложных программ, а также значительно выделяются авторы с большим опытом программирования на C/C++.

По результатам анализа вышеперечисленных источников было принято реше-

ние использовать подход, основанный на построении абстрактных синтаксических деревьев и классификации при помощи алгоритма Random Forest.

3 Моделирование

Описание процесса определения авторства исходного кода программ в виде модели «черного ящика» согласно методологии IDEF0 представлено на рисунке 3.1, его декомпозиция — на рисунке 3.2.

Рисунок 3.1 – Модель «черного ящика» процесса определения авторства исходного кода по методологии IDEF0

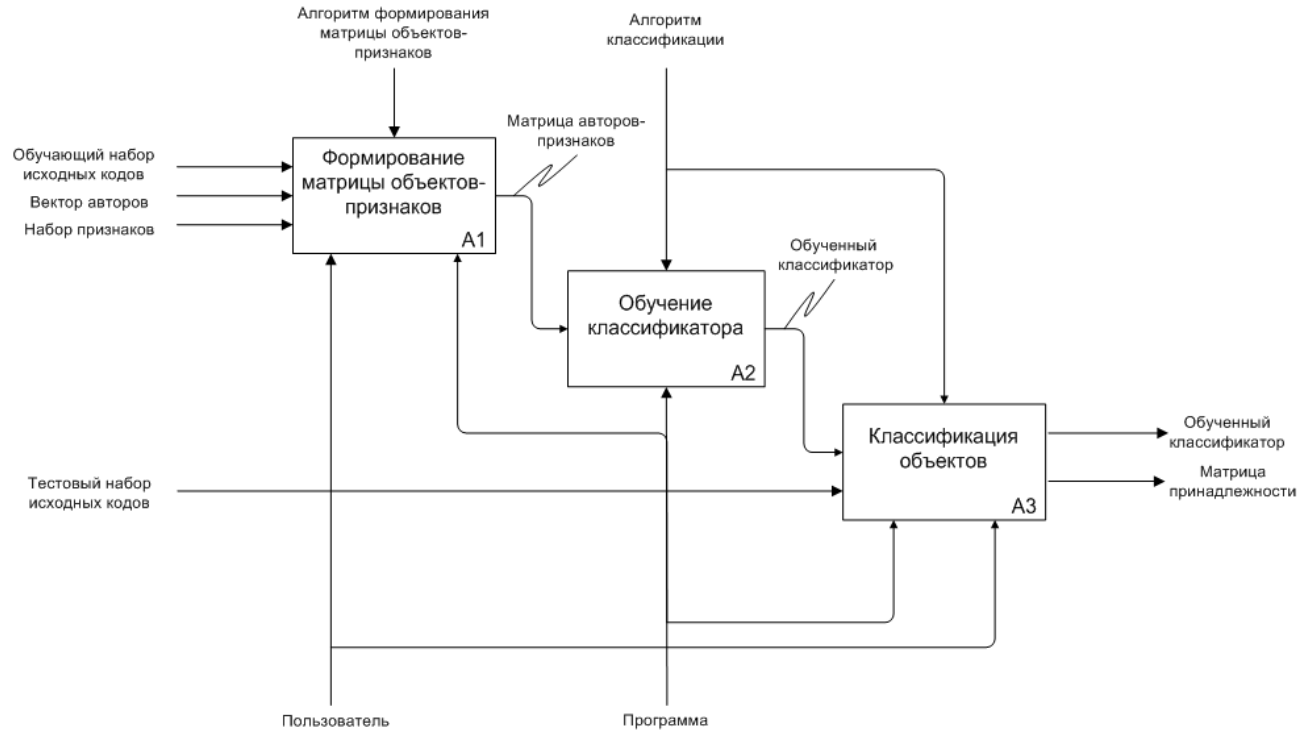

Рисунок 3.2 – Декомпозиция «черного ящика» процесса определения авторства исходного кода по методологии IDEF0

Заключение

По результатам, полученным в ходе научно-исследовательской работы, можно сделать следующие выводы:

1) Тестовые данные (см. раздел **??**) имели определенные недостатки, которые в конечном итоге привели к снижению точности классификации:

– программы, на которых производилось обучение и тестирование классификатора, были написаны студентами в ходе изучения основ программирования, т.е. у авторов в выборке отсутствовал опыт программирования на C/C++;

– среднее количество строк кода на файл составило всего 45 строк;

– в основном примеры содержали множество конструкций ввода-вывода и простые расчеты (например, площадей различных геометрических фигур);

– задания выполнялись по примерам из методического обеспечения, что повлияло на предпочтение использования тех или иных конструкций языка.

2) Не все признаки, используемые в классификации, равнозначны. Так, на-

пример, частоты ключевых слов «int» или «float» (см. раздел **??**) не так важны для определения стилистических особенностей написания программы, как, к примеру, определенные предпочтения автора при назначении идентификаторов.

3) Оптимальные параметры для задач классификации с помощью алгоритма Random Forest (см. раздел **??**) были подобраны эмпирически (см. раздел **??**) и соответствовали рекомендациям из различных источников.

В ходе работы был построен классификатор, позволяющий отнести исходный код к конктретному автору с точностью около 73%.

Среди задач на будущее можно выделить следующие:

– расширение набора признаков, а также расчет (с использованием экспертной оценки) веса для каждого из них (см. раздел **??**) для повышения точности классификации;

– подбор тестовых данных более сложных программ для большего количества авторов; при этом желательно, чтобы авторы программ имели некоторый опыт программирования на языке C/C++;

– исследование набора признаков для программ на других языках высокого уровня;

– исследование алгоритмов обнаружения плагиата в исходных кодах программ.

Список использованных источников

1 Образовательный стандарт ВУЗа ОС ТУСУР 01-2013 [Электронный ресурс]. — Режим доступа: https://storage.tusur.ru/files/40668/rules_tech_01-2013.pdf (дата обращения: 15.12.2016).

2 Source Code Authorship Analysis For Supporting The Cybercrime Investigation Process. Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis [Электронный ресурс]. — Режим доступа: http://www.icsd.aegean.gr/lecturers/stamatatos/papers/ICETE2005.pdf (дата обращения: 23.10.2016).

3 Маевский Д.А. Определение авторства программного обеспечения по исходному коду программ [Электронный ресурс]. — Режим доступа: http://www.khai.edu/csp/nauchportal/Arhiv/REKS/2014/REKS614/Maevsky.pdf (дата обращения: 17.12.2016).

4 Using Classification Techniques to Determine Source Code Authorship. Brian N. Pellin Computer Sciences Department University of Wisconsin [Электронный ресурс]. — Режим доступа: https://pdfs.semanticscholar.org/f9aa/790191a50bed02a877e1696c7bb71ea9f33a.pdf (дата обращения: 23.10.2016).

5 Source Code Authorship Attribution using n-grams. Steven Burrows, S.M.M. Tahaghoghi [Электронный ресурс]. — Режим доступа: https://pdfs.semanticscholar.org/79a2/1998c2f0afe2c616c01d590d6d0f6e16e9eb.pdf (дата обращения: 23.10.2016).

6 De-anonymizing Programmers via Code Stylometry. Aylin Caliskan-Islam, Drexel University; Richard Harang, U.S. Army Research Laboratory; Andrew Liu, University of Maryland; Arvind Narayanan, Princeton University; Clare Voss, U.S. Army Research Laboratory; Fabian Yamaguchi, University of Goettingen; Rachel Greenstadt, Drexel University [Электронный ресурс]. — Режим доступа:

https://www.usenix.org/system/files/conference/usenixsecurity15/sec15-paper-caliskan-islam.pdf (дата обращения: 07.09.2016).

7 Random Forest. Applied Multivariate Statistics — Spring 2012 [Электронный ресурс]. — Режим доступа: http://stat.ethz.ch/education/semesters/ss2012/ams/slides/v10.2.pdf (дата обращения: 23.10.2016).

Приложение А

(Обязательное)

Компакт-диск

Компакт-диск содержит:

– электронную версию пояснительной записки в форматах *.tex и *.pdf;

– актуальную версию программы, реализованную на языке программирования Python, для определения авторсва исходного кода программ на языке C/C++;

– тестовые данные для работы с программой.