МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

Тема работы

КУРСОВАЯ РАБОТА
по дисциплине «Программная инженерия»
ЮУрГУ – 09.03.04.2023.308-066.КР

Нормоконтролер, профессор
кафедры СП

_____ М.Л. Цымблер

“____”_____ 2024 г.

Научный руководитель
доктор физ.-мат. наук

_____ М.Л. Цымблер

Автор работы,
студент группы КЭ-303

_____ А.В. Толмачева

Работа защищена
с оценкой: _____

“____”_____ 2024 г.

Челябинск-2024

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

УТВЕРЖДАЮ
Зав. кафедрой СП
_____ Л.Б. Соколинский
“____”_____ 2024 г.

ЗАДАНИЕ
на выполнение выпускной курсовой работы
по дисциплине «Программная инженерия»
студенту группы КЭ-303
Толмачевой Анастасии Вячеславовне,
обучающемуся по направлению
09.03.04 «Программная инженерия»

1. **Тема работы**
   Разработка системы для выявления фиктивных аккаунтов Open Journal System.
2. **Срок сдачи студентом законченной работы:** 31.05.2024.
3. **Исходные данные к работе**
   3.1. Open Journal Systems. [Электронный ресурс]. URL: https://openjournalsystems.com/ (дата обращения 18.09.2023)
   3.2. Open Journal Systems. [Электронный ресурс]. URL:https://ru.wikipedia.org/wiki/Open_Journal_Systems (дата обращения 02.10.2023)
   3.3. Breuer A., Khosravani N., Tingley M., Cottel B. Preemptive Detection of Fake Accounts on Social Networks via Multi-Class Preferential Attachment Classifiers // KDD ’23: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023. – 105–116 pp. DOI: 10.1145/3580305.3599471
4. **Перечень подлежащих разработке вопросов**
   4.1. Анализ предметной области и литературы по теме работы.
   4.2. Разработка алгоритма выявления фиктивных аккаунтов.
   4.3. Проектирование интерфейса программной системы и модульной структуры приложения.
   4.4. Реализация программной системы, выявляющей фиктивные аккаунты, на основе разработанного алгоритма.
   4.5. Подготовка набора тестов и тестирование программной системы.
5. **Дата выдачи задания:** ____”_____ 2024 г.

Научный руководитель М.Л. Цымблер

Задание принял к исполнению А.В. Толмачева

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 5
1. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ 7
1.1. Описание предметной области 7
1.2. Обзор литературы 7
2. ПРОЕКТИРОВАНИЕ 9
2.1. Требования к системе 9
2.2. Варианты использования системы 9
2.3. Архитектура приложения 9
2.4. Графический интерфейс 9
2.5. Сравнение алгоритмов 9
3. РЕАЛИЗАЦИЯ 16
3.1. Программные средства реализации 16
3.2. Подготовка данных 16
3.3. Инженерия признаков 16
3.4. Реализация компонентов приложения 16
3.5. Реализация пользовательского интерфейса 16
4. ТЕСТИРОВАНИЕ 17
4.1. Функциональное тестирование 17
4.2. Вычислительные эксперименты 17
ЗАКЛЮЧЕНИЕ 18
ЛИТЕРАТУРА 19

ВВЕДЕНИЕ

Актуальность темы

Интернет может нести не только пользу, но также и вред. Одна из его опасностей - фиктивные аккаунты. Они существуют как в различных социальных сетях, так и на других платформах; некоторые используются в безобидных целях, а другие — для распространения ложной информации.

«Фейки» могут быть созданы с разными целями: получить коммерческую выгоду, дискредитировать настоящего пользователя, заполучить личную информацию и так далее [3]. Это очень серьезная опасность, которую не всегда можно распознать с первого взгляда. Некоторые пользователи мировой сети не догадываются, кто скрывается в диалоге с интернет-знакомым — люди могут умело подделывать аккаунты в социальных сетях. Кроме того, часто бывает, что за страницей обычных пользователей могут скрываться автоматизированные программы. Они могут практически не отличаться от обычных аккаунтов.

Особенно сильно это наносит вред такой среде, как наука. Подделка информации в этой сфере несет огромный вред, который может распространяться на все общество. Научные данные используются везде: от строительства домов, до лечения животных, и неточности или ошибки в них могут стоить дорого.

Основная проблема заключается в том, что фиктивные аккаунты, которые создаются в системах, связанных с наукой, тяжело отличимы от обычных аккаунтов. Для поддержания стабильной работы публикации статей необходимо быстрое реагирование на «фейки» и их оперативное удаление.

Цель и задачи

Целью курсовой работы является реализация программной системы, выявляющей фиктичные аккаунты в системе Open Journal Systems. Для достижения поставленной цели необходимо решить следующие задачи:

1. Провести анализ предметной области и литературы по теме работы.
2. Разработать алгоритм выявления фиктивных аккаунтов.
3. Спроектировать интерфейс программной системы и модульной

структуры приложения.

4. Реализовать программную систему, выявляющую фиктивные аккаунты, на основе разработанного алгоритма.

5. Подготовить набор тестов, выполнение тестирования программной системы.

Структура и содержание работы

Написать разделы, из которых состоит курсовая работа.

1. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

В разделе 1.1. располагается описание предметной области. В разделе 1.2. рассматриваются аналоги и существующие решения.

1.1. Описание предметной области

Open Journal System — это программное обеспечение, которое позволяет публиковать статьи и организовать рабочий процесс издательства. На ее основе разработаны многие порталы, работают институты, научные центры и журналы в разных странах мира (интерфейс OPS переведён более чем на 30 языков). Платформа обладает модульной структурой и имеет возможность подключения плагинов.

OJS может быть рассмотрена как электронная библиотека: программа обеспечивает доступ к контенту, поиск по нему (автора, ключевые слова, названия статей, год выпуска и так далее).

1.2. Обзор литературы

Для выбора наилучшего подхода к решению задач и достижения цели были прочитаны некоторые статьи на тему выявления «фейков». Исследования, посвящённые поиску фиктивных аккаунтов, представлены в следующих работах:

1. В работе «Fake Accounts Identification in Mobile Communication Networks Based on Machine Learning» [3] раскрывается проблема поддельных страниц. Их количество растет вместе с увеличением числа активных пользователей социальных сетей. Поддельные профили на сайтах социальных сетей создают ненастоящие новости и распространяют нежелательные материалы, содержащие спам-ссылки.

   В этой статье приводится контролируемый алгоритм машинного обучения, называемый машиной опорных векторов (SVM), который используется вместе с методом случайного леса. Эту концепцию можно легко применить для идентификации миллионов учетных записей, которые невозможно проверить вручную. Данная модель сравнивается с другими методами идентификации, и результаты

показывают, что предложенный алгоритм работает с большей точностью.

2. Статья «Social Networks Fake Profiles Detection Based on Account Setting and Activity» [1] посвящена выявлению поддельных профилей в социальных сетях. Подходы к выявлению поддельных профилей в социальных сетях можно разделить на подходы, направленные на анализ данных профилей. Для обнаружения поддельных профилей было предложено множество алгоритмов и методов. В этой статье оценивается влияние использования дерева решений (DT) и наивного алгоритма Байеса (NB) для классификации профилей пользователей на поддельные и подлинные.

3. В работе «Novel approaches to fake news and fake account detection in OSNs: user social engagement and visual content centric model» [4] исследователи используют модель SENAD, которая определяет подлинность новостных статей, публикуемых в Twitter, на основе подлинности и предвзятости пользователей, которые взаимодействуют с этими статьями. Предлагаемая модель включает в себя идею оценки соотношения подписчиков, возраст аккаунта и т.д. Для анализа изображений предлагается использовать нейронную сеть (CredNN). Предложенная гибридная идея объединения ELA и Sent и анализа настроений играет важную роль в обнаружении поддельных изображений с точностью около 76%.

4. В исследовании «Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach» [2] проведен анализ 62 миллионов общедоступных профилей пользователей Twitter, и разработана стратегия идентификации автоматически сгенерированных поддельных профилей. Используя алгоритм сопоставления шаблонов имен, анализ времени обновления твитов и времени создания профилей, было выявлено множество фиктивных учетных записей пользователей.

slashbox,pict2e

2. ПРОЕКТИРОВАНИЕ

В разделе 2.1. представлены функциональные и нефункциональные требования к системе. В разделе 2.2. описаны варианты использования системы. В разделе в разделе 2.3. представлена архитектура приложения, а в разделе 2.4. показан графический интерфейс, а в разделе 2.5. сравниваются различные алгоритмы выявления фиктивных аккаунтов.

2.1. Требования к системе

Функциональные требования

Функциональные требования определяют функциональность программного обеспечения, то есть описывают, какое поведение должна предоставлять разрабатываемое приложение

Нефункциональные требования

К нефункциональным требованиям системы относятся свойства, которыми она должна обладать. Например, удобство использования, безопасность, расширяемость и т.д.

2.2. Варианты использования системы

Описание способов взаимодействия с системой, кто с ней может работать, каким образом.

2.3. Архитектура приложения

Показать модули приложения, расписать то, что делает каждый из них.

2.4. Графический интерфейс

Представление графического интерфейса программной системы.

2.5. Сравнение алгоритмов

Для нахождения наилучшего алгоритма выявления фиктивных аккаунтов были разработаны пробные модели, а после найдены их критерии качества.

Первое, что было сделано - это извлечение признаков, по которым

проводилась кластеризация. Признаки:

- user_id: уникальный идентификатор пользователя. Используется для идентификации каждого аккаунта;
- username_length: длина имени пользователя;
- numbers_in_name: переменная, которая означает наличие цифр в имени пользователя. Если нет – ставится 0, иначе – 1;
- email_length: длина email;
- matching_names: переменная, которая означает совпадение username с email по определенному порогу сходства (в случае несовпадения 0, иначе – 1);
- pattern_email: продходит ли email по шаблону user@domain.com (в случае прохождения по парамтру 0, иначе – 1);
- country: проверка, указана ли страна (если не указана - 0, иначе - 1);
- date_last_email: проверка даты последнего отправленного email. Если она есть – 1, иначе – 0;
- date_registered: дата регистрации аккаунта;
- date_last_login: дата последнего входа в аккаунт;
- matching_dates: характеристика, совпадают ли даты послегнего входа в аккаунт и даты регистрации.

Далее подробнее рассмотрим алгоритмы кластеризации, которые были использованы для нахождения аномалий в виде фиктивных аккаунтов.

Алгоритм изолированного леса

Была произведена подготовка данных в виде заполнения пропущенных значений средним значением по столбцу. Была реализована нормализация признаков: они преобразованы в данные в диапазоне от 0 до 1.

Установлена ожидаемая доля аномалий - 5%. Аномалии (данные с индексом -1) были занесены в файл, а также создано графическое изображение для результатов кластеризации (рис. 1).

Рис. 1. Кластеризация методом изолированного леса

После этого были произведены расчеты метрик алгоритма, с использованием ячеек матрицы ошибок (табл. 1), где TP – число объектов, предсказанных моделью как положительные, которые действительно являются положительными, TN – число объектов, предсказанных моделью как отрицательные, которые действительно являются отрицательными, FP – число объектов, предсказанных моделью как положительные, которые на самом деле являются отрицательными, FN – число объектов, предсказанных моделью как отрицательные, которые на самом деле являются положительными.

Табл. 1. Прогноз алгоритма

True Positive	7
True Negative	151
False Positive	6
False Negative	90

Для оценки качества работы алгоритма необходимо ввести метрики accuracy (аккуратность), precision (точность) и recall (полнота). Первая показывает долю верно классифицированных объектов, вторая – долю объ-

ектов, которые модель классифицировала как положительные, и которые действительно являются положительными, а третья – долю объектов положительного класса, которые модель определила правильно. Рассчитаем их значения.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0,63$$

$$precision = \frac{TP}{TP + FP} = 0,62$$

$$recall = \frac{TP + TN}{TP + FN} = 0,08$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 0,14$$

Алгоритм иерархической кластеризации

Была произведена подготовка данных в виде заполнения пропущенных значений средним значением по столбцу. Была реализована нормализация признаков: они преобразованы в данные в диапазоне от 0 до 1. Была произведена сортировка по username_length. Создана матрица расстояний с методом Single Linkage, так как он сильнее реаширует на выбросы. Создано графическое изображение для результатов кластеризации (рис. 2).

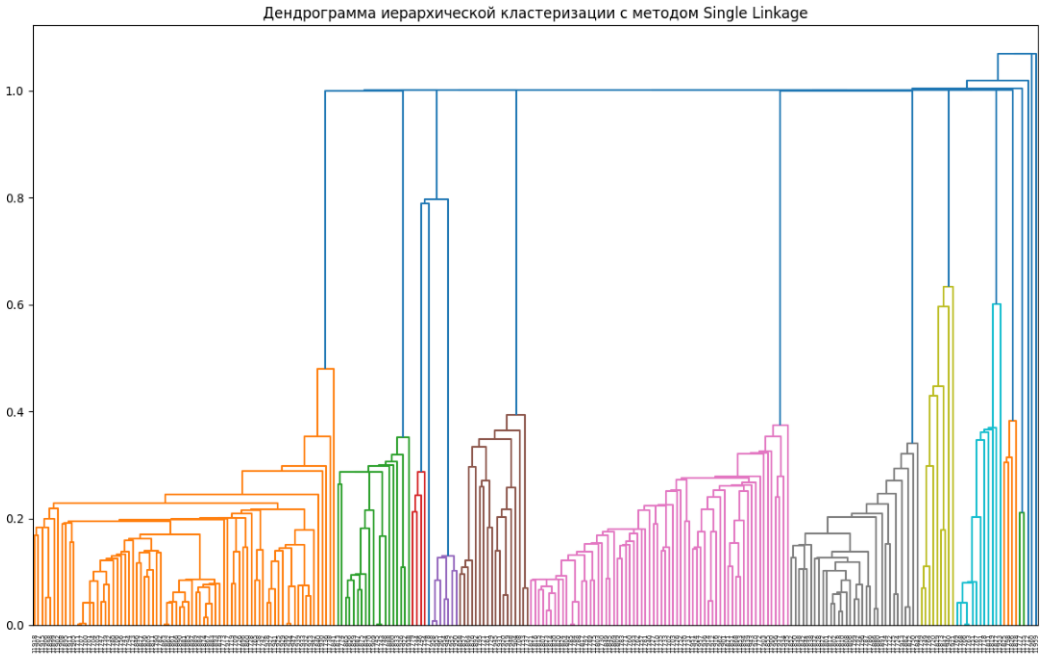

Рис. 2. Иерархическая кластеризация

После этого были произведены расчеты метрик алгоритма, с использованием ячеек матрицы ошибок (табл. 2).

Табл. 2. Прогноз алгоритма

True Positive	8
True Negative	152
False Positive	5
False Negative	89

Для оценки качества работы рассчитаем метрики accuracy (аккуратность), precision (точность) и recall (полнота).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0,60$$

$$precision = \frac{TP}{TP + FP} = 0,3$$

$$recall = \frac{TP + TN}{TP + FN} = 0,03$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 0,05$$

Алгоритм DBSCAN

Была произведена подготовка данных в виде заполнения пропущенных значений средним значением по столбцу. Была реализована нормализация признаков: они преобразованы в данные в диапазоне от 0 до 1. Кластеризация была произведена с параметрами eps = 1, min_samples = 6. Выбраны аномалии с индексом -1 и создано графическое изображение для результатов кластеризации (рис. 3).

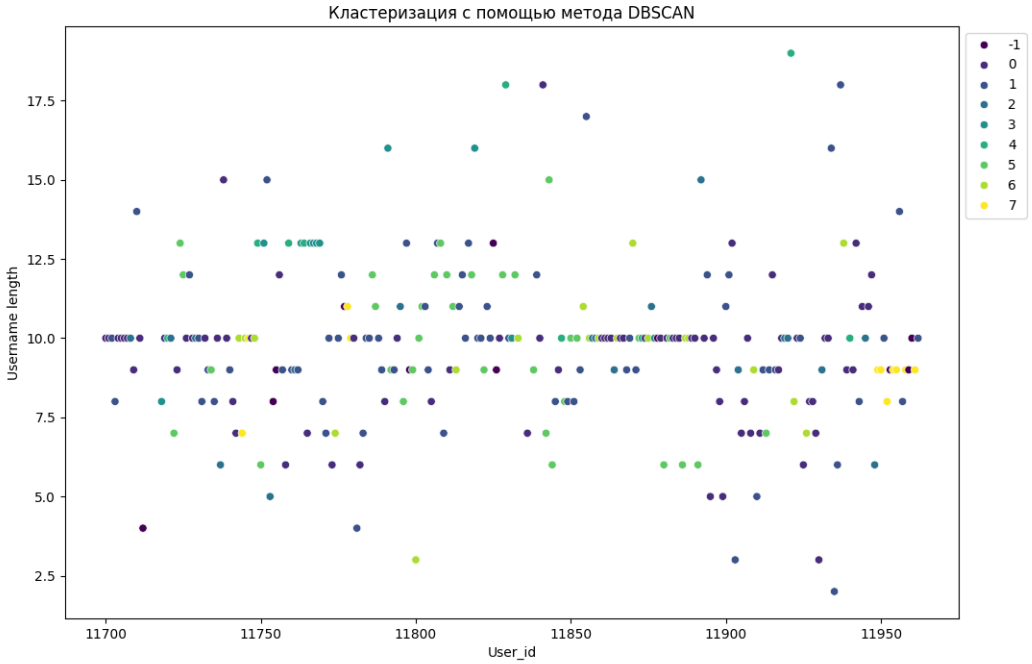

Рис. 3. Иерархическая кластеризация

После этого были произведены расчеты метрик алгоритма, с использованием ячеек матрицы ошибок (табл. 3).

Табл. 3. Прогноз алгоритма

True Positive	8
True Negative	154
False Positive	3
False Negative	89

Для оценки качества работы алгоритма необходимо рассчитать ввести метрики accuracy (аккуратность), precision (точность) и recall (полнота).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0,60$$

$$precision = \frac{TP}{TP + FP} = 0,3$$

$$recall = \frac{TP + TN}{TP + FN} = 0,03$$

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 0,05$$

Сравнение трех алгоритмов

Результаты рассчитанных метрик для всех реализованных алгоритмов представлены в таблице 4.

Табл. 4. Метрики алгоритмов

	accuracy	precision	recall	F
Isolation forest	0,63	0,62	0,08	0,14
Hierarchical clustering	0,60	0,3	0,03	0,05
DBSCAN	0,64	0,78	0,07	0,13

3. РЕАЛИЗАЦИЯ

3.1. Программные средства реализации

Прописать используемые языки программирования и библиотеки.

3.2. Подготовка данных

Сбор данных о пользователях системы электронного журнал, определение, какие аккаунты являются фиктивными.

3.3. Инженерия признаков

У фиктивных аккаунтов есть некоторые признаки, которые могут отличать их от обычных аккаунтов. Далее рассмотрим их подробнее.

1. Недавняя дата регистрации (по исследованию [4] 15.80% фиктивных аккаунтов и 2.80% настоящих аккаунтов имели регистрацию, совершенную в течение прошедшего месяца)
2. Дата последнего визита страницы (в основном страницу после создания не посещают)
3. Статус страницы (удалена, заблокирована)
4. Отсутствие логина или его части в адресе почты
5. Пользователь должен сменить пароль (что это за столбец)

3.4. Реализация компонентов приложения

Расписать работу модулей приложения, какие данные получают на входе, какие на выходе.

3.5. Реализация пользовательского интерфейса

Каким образом и где реализован интерфейс приложения. Описать основные механизмы и привести изображения.

4. ТЕСТИРОВАНИЕ

4.1. Функциональное тестирование

Провести тестирование на соответствие приложения предъявленным требованиям.

4.2. Вычислительные эксперименты

В данном разделе представлены вычислительные эксперименты для набора данных.

ЗАКЛЮЧЕНИЕ

ЛИТЕРАТУРА

1. Elyusufi Y., Elyusufi Z., Kbir M.A. Social networks fake profiles detection based on account setting and activity. // Proceedings of the 4th International Conference on Smart City Applications, SCA 2019, Casablanca, Morocco, October 02-04, 2019. – ACM, 2019. – P. 37:1–37:5. – URL: https://doi.org/10.1145/3368756.3369015.

2. Gurajala S. Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. / S. Gurajala, J.S. White, B. Hudson, J.N. Matthews. // Proceedings of the 2015 International Conference on Social Media & Society, Toronto, ON, Canada, July 27-29, 2015 / Ed. by A.A. Gruzd, J. Jacobson, P. Mai, B. Wellman. – ACM, 2015. – P. 9:1–9:7. – URL: https://doi.org/10.1145/2789187.2789206.

3. Hassan A., Alhalangy A.G.I., Al-Zahrani F. Fake Accounts Identification in Mobile Communication Networks Based on Machine Learning. // Int. J. Interact. Mob. Technol. – 2023. – Vol. 17. – No. 4. – P. 64–74. – URL: https://doi.org/10.3991/ijim.v17i04.37645.

4. Uppada S.K. Novel approaches to fake news and fake account detection in OSNs: user social engagement and visual content centric model. / S.K. Uppada, K. Manasa, B. Vidhathri, R. Harini, B. Sivaselvan. // Soc. Netw. Anal. Min. – 2022. – Vol. 12. – No. 1. – P. 52. – URL: https://doi.org/10.1007/s13278-022-00878-9.