Санкт-Петербургский Государственный Университет

Математико-механический факультет

Программная инженерия

Иванов Кирилл Андреевич

Учебная практика
(научно-исследовательская
работа)

Набор инструментов для анализа профилей пользователей: технологии веб-фреймворка Django

Научный руководитель:
к.т.н., доцент каф. информатики
Абрамов Максим Викторович

Санкт-Петербург
2021

Оглавление

ВВЕДЕНИЕ

Актуальность. Благодаря техническому прогрессу появляется возможность использовать в информационных системах все более совершенные методы хранения, передачи и защиты информации. Тем не менее наблюдается тенденция роста размера ущерба от киберпреступлений [10].

Информационные системы могут оказаться уязвимыми из-за недостаточной осведомленности в области информационной безопасности их операторов и пользователей. Так, злоумышленники могут получить доступ к интересующей их информации при помощи социоинженерных атак — прикладных психологических и аналитических приемов, которые применяются для скрытой мотивации пользователей публичной или корпоративной сети к нарушениям устоявшихся правил и политик в области информационной безопасности [9].

Атаки с применением социальной инженерии занимают лидирующие позиции среди киберпреступлений, о которых сообщалось в 2019 году [12]. Анализ возможности таких атак — это проблема, встающая перед специалистами в области информационной безопасности. Одним из решений является сбор информации о пользователях, сопряженный с анализом их уязвимых мест и оценкой возможных сценариев распространения атак внутри компаний. Такую информацию можно извлекать из страниц в социальных сетях.

В настоящее время «ВКонтакте» и «Одноклассники» являются одними из самых популярных социальных сетей в России [11], поэтому видится целесообразным собирать информацию на этих платформах.

Степень разработанности темы. По тематике защиты пользователей от социоинженерных атак существует ряд теоретических и практических

разработок. Так, исследователями лаборатории теоретических и междисциплинарных проблем информатики Санкт-Петербургского Федерального исследовательского центра РАН (ТиМПИ СПб ФИЦ РАН) был построен профиль уязвимостей пользователя в контексте социоинженерных атак [2, 3], включающий в себя оценку психологических особенностей [7, 8], информация для которой может быть получена из социальных сетей. Были предложены подходы к решению задачи определения вероятного пропущенного значения атрибута (город, возраст) профиля пользователя социальной сети [1], а также были получены результаты, позволяющие идентифицировать аккаунты пользователей в двух социальных сетях [5]. Тем не менее, существует необходимость разработки единого комплекса, агрегирующего результаты из перечисленных исследований.

Целью работы является увеличение оперативности процесса анализа профилей пользователей социальных сетей через создание автоматизированного исследовательско-практического комплекса.

Для достижения цели были поставлены следующие задачи:

- идентифицировать технологии и методы для получения, обработки и представления данных из социальных сетей;
- разработать архитектуру программного комплекса и его первый прототип;
- определить и внедрить методы и подходы для восстановления пропущенных атрибутов профилей социальных сетей, сопоставления профилей из разных социальных сетей.

Объектом исследования являются аккаунты пользователей в социальных сетях как источник данных об уязвимостях пользователей и их подверженности социоинженерным атакам.

Предметом исследования являются методы автоматизированного

извлечения, предобработки, унификации и представления данных со страниц пользователей социальных сетей.

Теоретическая и практическая значимость. Теоретическая значимость заключается в комбинировании созданных ранее методов и подходов для автоматизации процесса оценки защищенности пользователей соцсетей от социоинженерных атак. Практическая значимость заключается в разработке веб-приложения, позволяющего анализировать аккаунты пользователей социальных сетей с целью выявления психологических особенностей и уязвимых мест, что необходимо для построения моделей распространения социоинженерных атак. Такая система будет полезна как исследователям социальных сетей, так и коммерческим организациям, желающим своевременно предпринять превентивные меры по предотвращению социоинженерных атак.

Методология учебной практики заключается в постановке и формализации задач, связанных с автоматизированным извлечением данных из социальных сетей, описанием моделей и сущностей, используемых для преобразования, хранения и унификации представления данных, а также реализация прототипа программного комплекса.

Методы. Для реализации практической части работы использовались методы объектно-ориентированного программирования. Программная реализация осуществлялась на языке программирования Python. В качестве вспомогательных технических средств были выбраны Django, PostgreSQL, Bootstrap 4.

Структура работы. Текст данной работы состоит из введения, трех глав, заключения и списка используемой литературы. Общий объем — 22 страницы.

В 1 главе обосновывается выбор и постановка цели и задач.

Во 2 главе описываются известные наработки и алгоритмы на тему анализа социальных сетей и различные средства, послужившие основой для проведения данной работы. Также в этой главе проводится выбор используемых технических средств.

В 3 главе приведено описание реализованного прототипа веб-приложения и его модулей.

1 ОБОСНОВАНИЕ ЦЕЛЕЙ И ЗАДАЧ

Целью работы является увеличение оперативности процесса анализа профилей пользователей социальных сетей через создание автоматизированного исследовательско-практического комплекса. Практические наработки по тематике социоинженерных атак зачастую разрознены, что затрудняет полностью автоматизировать процесс анализа защищенности пользователя социальной сети, поэтому цель данной работы видится актуальной.

Для достижения цели были поставлены следующие задачи:

- идентифицировать технологии и методы для получения, обработки и представления данных из социальных сетей;
- разработать архитектуру программного комплекса и его первый прототип;
- определить и внедрить методы и подходы для восстановления пропущенных атрибутов профилей социальных сетей, сопоставления профилей из разных социальных сетей.

2 ОБЗОР ИСПОЛЬЗУЕМЫХ ПОДХОДОВ И РЕШЕНИЙ

В данной главе рассматриваются известные наработки и алгоритмы на тему анализа социальных сетей: методы экстракции данных из социальных сетей, подходы к восстановлению фрагментов метапрофиля пользователя, метод сопоставления профилей социальных сетей. Также определяются технические средства для разработки системы.

2.1 Экстракция данных из социальных сетей

Получить данные о пользователе социальной сети можно путем выгрузки HTML страницы пользователя с последующим выделение нужной информации — такой подход требует написания отдельного парсера для каждой социальной сети.

Альтернативно можно воспользоваться методами, которые предоставляет API социальной сети (API ВКонтакте [13], API Одноклассники [14]).

Нас интересует последний подход, поскольку с его помощью можно быстро получить информацию о большом количестве пользователей, что важно при анализе аккаунтов сотрудников крупных компаний.

2.2 Подходы к восстановлению фрагментов метапрофиля

Метапрофиль пользователя – это набор анкетных данных данного пользователя. Часто некоторые атрибуты метапрофиля бывают незаполненными или искаженными, что усложняет дальнейший анализ, поэтому появляется необходимость восполнения недостающих значений.

Уже существуют исследования, посвященные теме восстановления значений метапрофиля. Так, в работе [6] приводится классификация подходов к восстановлению. В публикации [4] содержится сравнение

статистического метода, метода на основе нейронных сетей и метода на основе кластеризации на социальные группы в задаче определения возраста пользователя. В целом, можно выделить следующие подходы к восстановлению.

Анализ аккаунтов в других соцсетях. Обязательные для заполнения поля в разных соцсетях могут отличаться, также пользователь может не указать какую-либо информацию о себе в одной соцсети, но указать в другой. Для восстановления фрагментов метапрофиля нужно установить принадлежность аккаунтов в разных соцсетях одному пользователю и создать единый профиль, который будет агрегировать информацию из этих соцсетей.

Анализ профиля в социальной сети. Этот подход основан на анализе информации, указанной на странице пользователя в социальной сети: при помощи некоторых алгоритмов и опубликованной информации делается предположение о недостающем атрибуте.

Анализ социального окружения. Данный подход основывается на анализе страниц друзей пользователя в соцсети: осуществляется поиск значений соответствующих атрибутов в профилях друзей, строится их статистическое распределение и делается предположение о недостающем атрибуте пользователя. В работе [1] предложены алгоритмы восстановления города и возраста путем анализа социального окружения.

Комбинированный подход. Используются комбинации описанных ранее подходов. Например, для восстановления возраста пользователя можно сделать два предположения: на основе даты окончания школы и на основе моды возраста социального окружения. После этого нужно выбрать наиболее вероятное предположение.

2.3 Сопоставление профилей социальных сетей

Задача сопоставления двух профилей в социальных сетях заключается в определении вероятности их принадлежности одному пользователю. Решение данной задачи необходимо для создания единого профиля пользователя, содержащего информацию из разных социальных сетей, что позволит предоставить большее количество данных для дальнейшего анализа.

Предлагаемый в исследовании [5] алгоритм сводит задачу сопоставления профилей в социальных сетях к задаче бинарной классификации. Для определения вероятности используются анкетные данные пользователя в социальной сети и данные социального окружения.

2.4 Технологии и инструменты

Python [15] был выбран в качестве языка программирования, на котором будет разработан прототип системы, потому что он имеет богатый набор библиотек для анализа данных, например, NumPy, pandas.

Django – это веб-фреймворк для языка Python, который способствует быстрой разработке и чистому, прагматичному дизайну [15]. Данный фреймворк следует архитектурному шаблону MVT (похожему на MVC), что способствует легкому масштабированию системы при добавлении обработки информации из других социальных сетей и подключении сервисов, выполняющих анализ и прогнозирование атак. Также Django предоставляет большой набор стандартных приложений (auth, sessions и др.) в рамках пакета contrib, что освобождает от поиска аналогичных решений и проверки их совместимости, — это важно при разработке прототипа системы.

PostgreSQL – это объектно-реляционная система баз данных с открытым исходным кодом, активная разработка которой насчитывает более 30 лет [17]. Данная СУБД имеет официальную поддержку Django и PL/Python — процедурного языка для описания функций Postgres.

Bootstrap – фреймворк, предоставляющий широкий набор инструментов для быстрого проектирования и создания веб-приложения на основе HTML и CSS [18]. Поскольку создается прототип веб-приложения, использование данного фреймворка целесообразно.

2.5 Вывод

В данной главе были рассмотрены методы и инструменты для получения данных о пользователях «ВКонтакте» и «Одноклассники», рассмотрены подходы к восстановлению фрагментов метапрофиля пользователя социальной сети, рассмотрены подходы к сопоставлению профилей пользователей социальных сетей и произведен выбор технических средств и инструментов.

3 РЕАЛИЗАЦИЯ ПРОТОТИПА

Данная глава посвящена описанию реализации прототипа целевого веб-приложения: описана его архитектура, основные модули и пользовательский интерфейс.

3.1 Архитектура

Прототип комплекса создан на языке Python с использованием веб-фреймворка Django и фреймворка Bootstrap. Ниже приведена диаграмма пакетов (рис. 3.1).

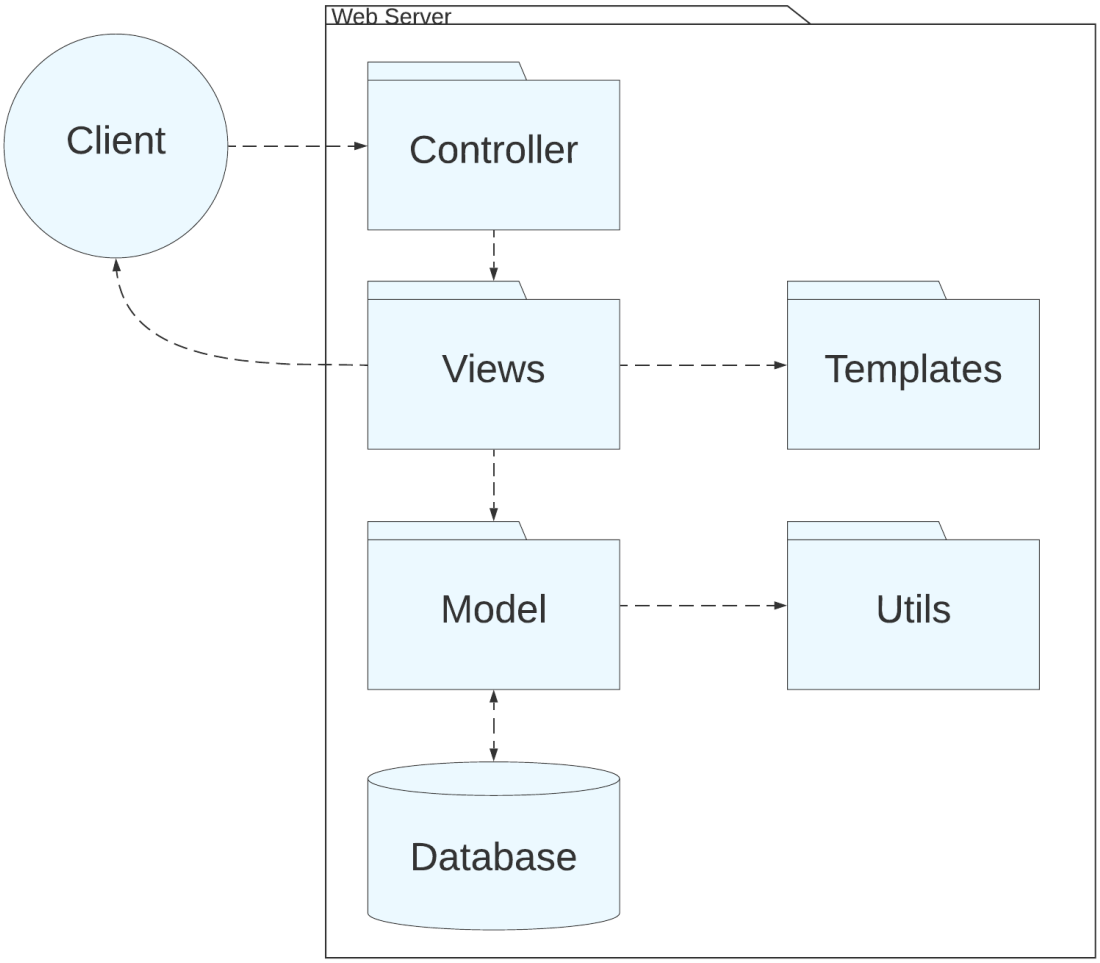

Рис. 3.1 — Диаграмма пакетов веб-приложения

Рассмотрим подробнее компоненты диаграммы.

- **Controller** представляет собой диспетчер адресов веб-страниц, использующихся в приложении. Данный пакет отвечает за выбор нужного обработчика адреса и передачу запроса этому обработчику.
- **Views** представляет собой набор функций, обрабатывающих запрос от контроллера. Views отвечает за отображение веб-страниц, формируя их на основе запроса, классов модели и HTML шаблонов.
- **Model** содержит классы, описывающие информацию в базе данных, а также логику заполнения атрибутов профиля по короткой ссылке или идентификатору «ВКонтакте».
- **Utils** содержит логику взаимодействия модели с API социальных сетей, а также код, отвечающий за восстановление фрагментов метапрофиля, визуализацию социального графа и сопоставление профилей в социальных сетях «ВКонтакте» и «Одноклассники».
- **Templates** содержит HTML-шаблоны страниц сервиса, написанные с применение шаблонизатора Django.
- **Database** – база данных, хранящая информацию о пользователях социальных сетей с восстановленными атрибутами.

3.2 Работа с данными социальных сетей

Работа с данными в прототипе веб-приложении состоит из следующих этапов.

Извлечение данных из социальных сетей. Для экстракции данных пользователей из социальных сетей «ВКонтакте» и «Одноклассники» была написана библиотеке на языке Python, работающая на основе HTTP-запросов к VK API и OK API. Для запроса используется

уникальный токен пользователя.

Восстановление метапрофиля. Для восстановления фрагментов метапрофиля был выбран алгоритм описанный в [1], что обусловлено простотой его реализации в прототипе веб-приложения. В дальнейшем планируется использовать алгоритмы, проводящие более детальный анализ пользователя, его окружения и цифровых следов.

Хранение. Информация из социальной сети преобразуется в объект модели Django, который описывает пользователя социальной сети, и помещается в базу данных на время анализа.

3.3 Сопоставление аккаунтов социальных сетей

Для определения вероятности принадлежности аккаунтов в социальных сетях «ВКонтакте» и «Одноклассники» одному пользователю был использован бинарный классификатор на основе логистической регрессии, описанный в [5] (коэффициенты регрессии были заранее подсчитаны). Для сопоставления профилей используются следующие атрибуты: «имя», «фамилия», «друзья», «город проживания», «дата рождения». В качестве предикторов логистической функции используются значения, полученные после сопоставления соответствующих атрибутов исходных профилей.

Значения атрибутов «имя», «фамилия» и «город проживания» подвергались предобработке: для атрибута «имя» – определение полного имени пользователя по его краткой форме при помощи сервиса Dadata [19] и приведение полученного значения к нижнему регистру, для атрибутов «фамилия» и «город проживания» – удаление специальных символов, транслитерация и приведение к нижнему регистру.

Вычисление значений предикторов производилось в соответствии с предложенными в [5] методами.

3.4 Визуализация социального графа

Социальный граф друзей пользователя – это тройка $G = \{u_0, F, L\}$, где u_0 – целевой пользователь, для которого строится социальный граф, F – друзья пользователя u_0 в данной социальной сети – вершины графа, $L = \{\{f_1, f_2\}: f_1, f_2 \in F, f_1$ и f_2 – друзья в соцсети$\} \cup \{\{u_0, f\}: f \in F\}$ – ребра графа.

Для построения социального графа для пользователя u_0 из социальных сетей извлекаются данные о его друзьях и формируются узлы F, содержащие идентификатор друга в социальной сети и его имя. После этого для каждого узла $f \in F$ находится список M общих с u_0 друзей, и в L добавляются неупорядоченные пары $\{f, u\}$ для всех $u \in M$.

На основе полученных данных строится визуализация при помощи библиотеки vis.js [20].

3.5 Пользовательский интерфейс

Пользовательский интерфейс написан на HTML с применением CSS и Bootstrap 4. Рассмотрим детальнее интерфейс пользователя.

Авторизация. Для входа в систему использовалась библиотека allauth, поддерживающая авторизацию и регистрацию с помощью различных соцсетей. Пользователь может зарегистрироваться в системе или авторизоваться при помощи социальных сетей «ВКонтакте» или «Одноклассники».

Далее приведена начальная страница сервиса. Пользователь должен авторизоваться, чтобы использовать функциональность приложения.

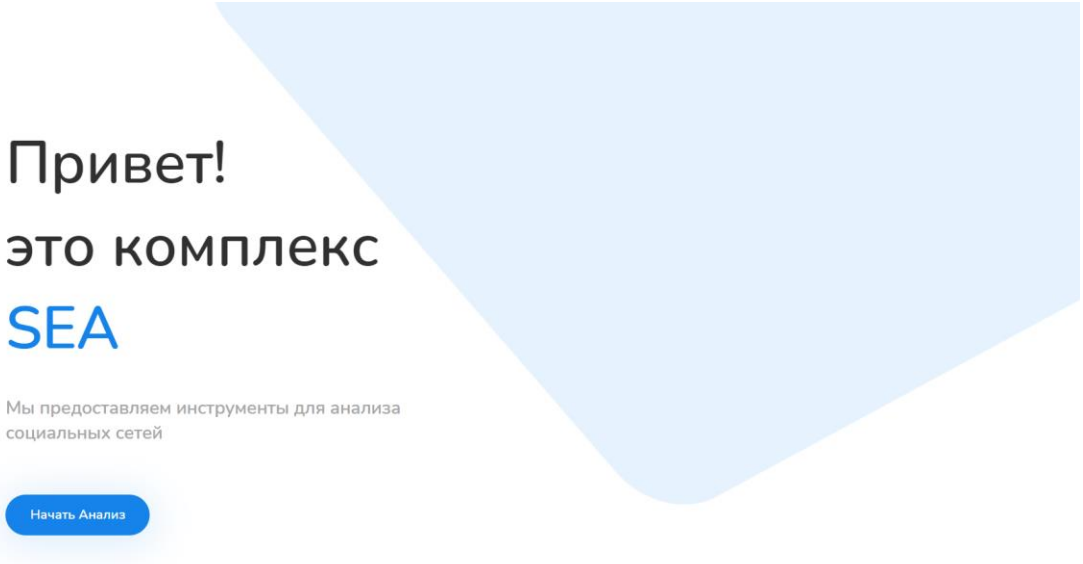

Рис. 3.2 — Стартовая страница

Поиск пользователя. Страница поиска пользователей содержит форму для ввода идентификатора. После отправки формы происходит перенаправление на страницу с результатами поиска (рис. 3.3), которая содержит интерактивный социальный граф друзей пользователя (рис. 3.4), анкетные данные в социальной сети, а также восстановленные значения. Ближайшие друзья пользователя определяются как пользователи, имеющие наибольшее количество связей в социальном графе.

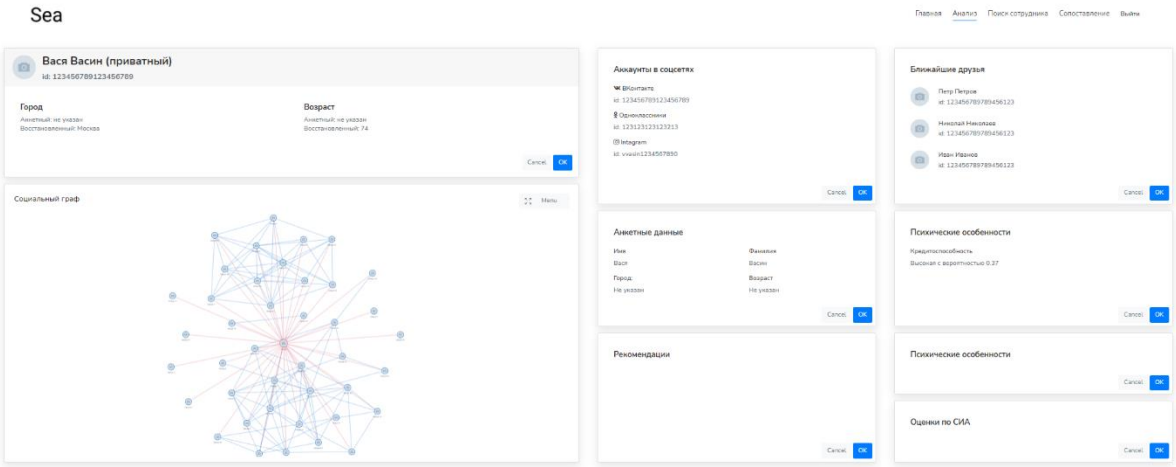

Рис. 3.3 — Страница с результатом поиска

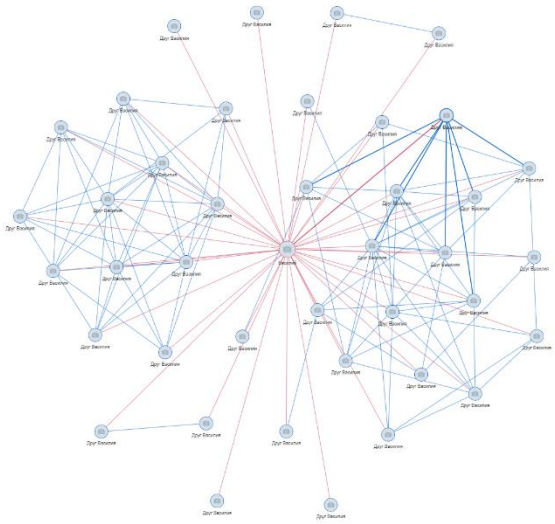

Рис. 3.4 — Визуализация социального графа

Сопоставление аккаунтов. Страница сопоставления аккаунтов «ВКонтакте» и «Одноклассники» содержит поля для ввода идентификатора пользователя в данных соцсетях (рис. 3.5). После отправки формы пользователь получает анкетную информацию об аккаунтах данных социальных сетей и процент их принадлежности одному пользователю.

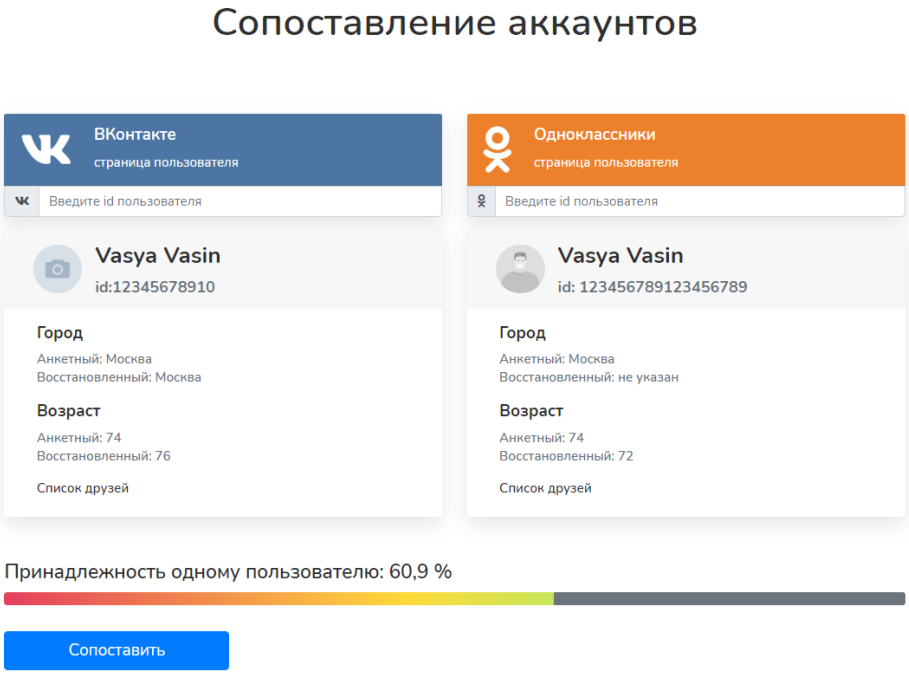

Рис. 3.5 — Страница сопоставления аккаунтов

При нажатии на кнопку «Список друзей» происходит открытие всплывающего окна, содержащего список друзей пользователя в данной социальной сети (рис. 3.6).

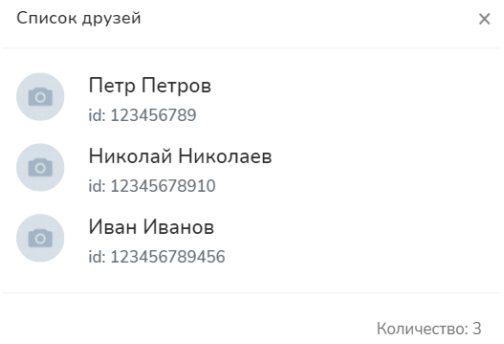

Рис. 3.6 — Окно со списком друзей

Если пользователь не указал дату рождения или город, произойдет восстановление соответствующего атрибута.

3.6 Вывод

В данной главе рассмотрена реализация прототипа веб-приложения и описана основная функциональность системы: экстракция данных из социальных сетей, восстановление значений атрибутов метапрофиля, сопоставление аккаунтов и визуализация социального графа.

ЗАКЛЮЧЕНИЕ

В ходе данной работы были выполнены следующие задачи:

- идентифицированы технологии и методы для получения, обработки и представления данных из социальных сетей;
- разработаны архитектура программного комплекса и его первый прототип;
- определены и внедрены методы и подходы для восстановления пропущенных атрибутов профилей социальных сетей, сопоставления профилей из разных социальных сетей.

Результатом данной работы является прототип веб-приложения, который может использоваться в исследованиях для сбора датасета из социальных сетей, анализа социальных связей и моделирования траекторий распространения социоинженерных атак.

Перспективы данного исследования заключаются в добавлении функциональности для работы с многоходовыми социоинженерными атаками, внедрении модели оценки выраженности психологических особенностей пользователей по их постам, а также реализации более продвинутых методов восстановления атрибутов пользователя.

СПИСОК ЛИТЕРАТУРЫ

[1] Абрамов М. В., Тулупьев А. Л., Тулупьева Т. В. Агрегирование данных из социальных сетей для восстановления фрагмента мета-профиля пользователя // Шестнадцатая Национальная конференция по искусственному интеллекту с международным участием КИИ-2018 Труды конференции: в 2-х томах. 2018. С. 189–197.

[2] Абрамов М.В., Тулупьев А.Л., Тулупьева Т.В. Психологические особенности, психические состояния пользователя и профиль его уязвимостей в контексте социоинженерных атак // Психология психических состояний: сб. статей студентов, магистрантов, аспирантов и молодых ученых. Казань. 2019. С. 312–317. ISBN 978-5-00130-159-2.

[3] Багрецов Г.И., Шиндарев Н.А., Абрамов М.В., Тулупьева Т.В. Подходы к разработке моделей для анализа текстовой информации в профилях социальной сети в целях построения профиля уязвимостей пользователя. // Сборник докладов Международной конференции по мягким вычислениям и измерениям (SCM-2017). Санкт-Петербург. Том 1-2. Т. 1. 2017. С. 134–137.

[4] Грезин В. С., Новосядлый В. А. О проблеме определения возраста участника социальной сети // Известия высших учебных заведений. Северо-Кавказский регион. Естественные науки. 2015.

[5] Корепанова А. А., Олисеенко В. Д., Абрамов М. В., Тулупьев А. Л. Применение методов машинного обучения в задаче идентификации аккаунтов пользователя в двух социальных сетях // Компьютерные инструменты в образовании. 2019. № 3. С. 29–43. doi:10.32603/2071-2340-2019-3-29-43.

[6] Соколова Т. В., Чеповский А. М. Проблема восстановления профилей пользователей социальных сетей // Вопросы кибербезопасности. 2019.

[7] Тулупьева Т.В., Тафинцева А.С., Тулупьев А.Л. Подход к анализу отражения особенностей личности в цифровых следах // Вестн. психотерапии. 2016. № 60 (65). С. 124–137.

[8] Тулупьева Т.В., Суворова А.В., Азаров А.А., Тулупьев А.Л., Бордовская Н.В. Возможности и опыт применения компьютерных инструментов в анализе цифровых следов студентов пользователей социальной сети // Компьютерные инструменты в образовании. 2015. № 5. С. 3–13.

[9] A. A. Azarov, T. V. Tulupyeva, A. V. Suvorova, A. L. Tulupyev, M. V. Abramov, and R. M. Jusupov, Socioengineering attacks. Problems of analysis, St Petersburg: Nauka Publ., 2016 (in Russian).

[10] Amount of monetary damage caused by reported cyber crime to the IC3 from 2001 to 2019. — URL:https://www.statista.com/statistics/267132/total-damage-caused-by-by-cyber-crime-in-the-us/.

[11] Most used social media platforms in Russia as of 3rd quarter 2019, by penetration rate. — URL: https://www.statista.com/statistics/284447/russia-social-network-penetration/.

[12] Types of cyber crime most frequently reported to the IC3 in 2019, by victim count. — URL: https://www.statista.com/statistics/184083/commonly-reported-types-of-cyber-crime/.

[13] ВКонтакте — Разработчикам. — URL: https://vk.com/dev (дата обращения: 25.09.2020).

[14] API Одноклассники. — URL: https://apiok.ru/ (дата обращения: 24.02.2021).

[15] Python Programming Language. — URL: https://www.python.org/ (дата обращения: 20.09.2020).

[16] Django: The Web framework for perfectionists with deadlines. — URL: https://www.djangoproject.com/ (дата обращения: 20.09.2020).

[17] PostgreSQL: The World's Most Advanced Open Source Relational Database. — URL: https://www.postgresql.org/ (дата обращения: 20.10.2020).

[18] Bootstrap: the world's most popular front-end open source toolkit. — URL: https://getbootstrap.com/ (дата обращения: 14.10.2020).

[19] Dadata. — URL: https://dadata.ru/ (дата обращения: 19.03.2021).

[20] Vis.js: a dynamic, browser based visualization library. — URL: https://visjs.org/ (дата обращения: 21.03.2021).

[21] django-allauth: integrated set of Django applications addressing authentication, registration, account management as well as 3rd party (social) account authentication. — URL: https://django-allauth.readthedocs.io/en/latest/ (дата обращения: 27.09.2021).