Агрегация данных из социальных сетей для определения наиболее вероятной конфигурации пропущенных значений параметров мета-профиля пользователя

М. В. Абрамов[1], Н. Е. Слезкин[2], Т. В. Тулупьева[3]
Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук
Санкт-Петербургский государственный университет
[1]mva16@list.ru, [2]ne.slezkin@gmail.com, [3]tvt100a@mail.ru

***Аннотация*. В данной работе представлен подход к решению задачи выбора наиболее вероятной конфигурации пропущенных значений мета-профиля пользователя, таких как родной город, год рождения, город проживания и др. Для решения этой задачи агрегируются данные из социальных сетей, получаемые из альтернативных источников – социального окружения пользователя, его анкетных данных, данных из его аккаунтов в других социальных сетях. Рассмотрены различные комбинации подходов к выбору гипотез о значениях пропущенных параметров мета-профиля пользователя.**

Ключевые слова: информационная безопасность; социоинженерные атаки; защита пользователя; профиль уязвимостей пользователя; мета-профиль пользователя; киберсоциальные системы; анализ защищённости

I. Введение

В настоящее время информационные технологии являются неотъемлемой частью жизни практически каждого человека. Информационные системы, хранящие критичные документы, окружают нас не только на работе, но и дома. Вместе с тем повышается актуальность вопросов информационной безопасности [5, 10, 12]. Выделяют программно-технические и социоинженерные атаки на информационную безопасность [9]. Вопросы анализа защищённости и защиты информационных систем от программно-технических атак достаточно хорошо изучены, существуют разработки, позволяющие снижать риск успеха таких атак [1, 3]. Вопросы, связанные с социоинженерными атаками, изучены меньше, хотя, согласно статистике [6], они происходят всё чаще и являются одними из наиболее эффективных. Вместе с тем, одной из важных задач в этой области является анализ защищённости пользователей информационных систем от таких атак. Автоматизированные системы анализа защищённости пользователей от социоинженерных атак позволили бы лицам, принимающим решения, осуществлять своевременные меры по повышению уровня безопасности. Общая цель направления исследований заключается в разработке таких систем, а также систем упреждающей диагностики и бэктрекинга инцидентов, рекомендательных систем.

В рамках этой ветки исследований необходимо решить ряд частных задач. В настоящее время профиль уязвимостей, на котором основаны оценки защищённости сотрудников компании, строится на базе некоторых особенностей личности пользователя [8]. Источником этих особенностей могут выступать знания экспертов, выбор пользователем стратегии в некоторой компьютерной игре, его аккаунт в социальной сети и другие. Данные, извлекаемые из аккаунта пользователя в одной социальной сети, могут быть недостаточно подробными. Для расширения объёма анализируемых данных предлагается использовать информацию, извлекаемую из аккаунта пользователя в других социальных сетях. Таким образом актуальной является задача поиска и идентификации аккаунтов одного пользователя в разных социальных сетях. Часто информация, содержащаяся в разных аккаунтах одного пользователя, отличается. Предлагается строить унифицированные мета-профили сотрудников компании на основании информации, извлекаемой из социальных сетей, а после сравнивать мета-профили.

Исследование посвящено решению актуальной задачи автоматизированного построения мета-профилей сотрудников компании на основании агрегации сведений, извлекаемых из социальных сетей. Мета-профиль пользователя представляет собой набор анкетных данных, таких как возраст, родной город, город проживания и иных. Часто информация, которая содержится в анкете пользователя на странице, не является достоверной, а иногда и не заполнена вовсе. Ранее был предложен подход к агрегации информации из социальных сетей о параметрах мета-профиля [7, 11]. В результате агрегации имеем несколько предположений о пропущенных значениях мета-профиля. Данная статья посвящена подходу к выбору совокупности данных, составляющих

Работа выполнена в рамках проекта по государственному заданию СПИИРАН № 0073-2018-0001, при финансовой поддержке РФФИ, проект №18-37-00323; проект №18-01-00626

наиболее вероятную конфигурацию набора параметров мета-профиля пользователя.

II. Релевантные работы

Данное исследование является частью общего исследования, направленного на повышение уровня защищённости пользователей информационных систем за счёт разработки автоматизированных средств анализа защищённости киберсоциальных систем от социоинженерных атак, проводимых на базе лаборатории теоретических и междисциплинарных проблем информатики Санкт-Петербургского института информатики и автоматизации РАН (ТиМПИ СПИИРАН) [9].

Вопросы идентификации разных аккаунтов пользователя в разных социальных сетях рассматривались в [9]. В данном исследовании идентифицировались аккаунты одного пользователя в социальных сетях ВКонтакте, Facebook и Instagram, в [7, 11] предложены подходы к восстановлению мета-профилей пользователей, построены алгоритмы определения пропущенных атрибутов мета-профиля. Например, для определения возраста пользователя агрегируется информация из его анкеты, из анкет в аккаунтах друзей и т.п. Также в этой работе представлены алгоритмы определения родного города пользователя и текущего города проживания.

Также ранее были предложены используемые в данной работе методы для поиска и идентификации аккаунтов сотрудников компании в социальной сети ВКонтакте [5]. Подходы к автоматизированному анализу степени выраженности некоторых особенностей личности пользователя на основании публикуемого контента в социальной сети [2].

III. Постановка общей задачи

Формализация задачи агрегации данных из социальных сетей для определения наиболее вероятной конфигурации пропущенных значений параметров мета-профиля пользователя может быть представлена следующим образом. Зададим матрицу, в которой в столбцах содержатся источники информации (социальное окружение пользователя – его друзья) о наборе параметров мета-профиля пользователя, например, о годе рождения, родном городе или ином. В строках – значения параметров мета-профиля, полученные из аккаунтов его социального окружения. Таким образом имеем следующую матрицу.

$$\begin{array}{c} \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{array} \begin{array}{c} \begin{array}{cccc} U_1 & U_2 & \cdots & U_n \end{array} \\ \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mn} \end{pmatrix} \end{array}$$

$U_1 \ldots U_n$ – пользователи из социального окружения анализируемого пользователя. Атрибуты $a_1 \ldots a_m$ – это элементы мета-профиля пользователя, например, пол, возраст, город проживания, родной город, религиозные взгляды и другие. Элементы матрицы S_{ij} – значения параметров мета-профилей пользователей, полученные из их аккаунтов в социальной сети. Отметим, что не во всех аккаунтах пользователей указаны все параметры. В случае пропуска значение элемента матрицы равно `0`.

Для иллюстрации рассмотрим следующий пример. Пусть у пользователя есть три друга и их анкеты на страницах в социальной сети содержат следующую информацию: пол, возраст и текущий город проживания. Представим эту информацию в виде соответствующей матрицы, где в столбцах указываются различные источники информации о параметрах мета-профиля – пользователи, друзья целевого, в строках – атрибуты мета-профиля, на основании которых строится мета-профиль целевого пользователя.

$$\begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} U_1 & U_2 & U_3 \end{array} \\ \begin{pmatrix} \text{м} & \text{м} & \text{ж} \\ 21 & 23 & 23 \\ \text{СПб} & \text{Мск} & \text{СПб} \end{pmatrix} \end{array}$$

Помимо информации, получаемой из анкетных данных аккаунтов друзей, предполагается использование также оценочной информации. Т.е. матрицу можно расширить данными, которые указал сам пользователь, а также информацией, которую можно извлечь из репостов пользователя, его групп и лайков, поставленных пользователем. Таким образом получим матрицу

$$\begin{array}{c} \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{array} \begin{array}{c} \begin{array}{ccccccccc} U_1 & U_2 & \cdots & U_n & \hat{\omega} & \omega^0 & \omega^1 & \cdots & \omega^k \end{array} \\ \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1n} & \hat{\omega}_1 & \omega_1^0 & \omega_1^1 & \cdots & \omega_1^k \\ S_{21} & S_{22} & \cdots & S_{2n} & \hat{\omega}_2 & \omega_2^0 & \omega_2^1 & \cdots & \omega_2^k \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mn} & \hat{\omega}_m & \omega_m^0 & \omega_m^1 & \cdots & \omega_m^k \end{pmatrix} \end{array},$$

где $\hat{\omega}$ – столбец, описывающий результат агрегации параметров мета-профилей пользователей из социального окружения исследуемого: на основании значений строки атрибута, определяется значение, которое встречается большее количество раз. Значение $\hat{\omega}$ всех трех атрибутов строится по выделению моды вектора. На примере атрибута возраста можно увидеть, что значение $\hat{\omega}$ равно 23, так как мода вектора $\{21, 23, 23\}$ равна 23. ω^0 – вектор атрибутов, полученных с личной страницы пользователя. Столбец ω^0 состоит из той информации, что сам пользователь указал на своей личной странице. Например, если пользователь указал 1996 год рождения, то в соответствующую ячейку матрицы поместиться значение 21. Если пользователь ничего не указал в личной анкете, то поле будет отсутствовать, как, например, пол или город проживания. $\omega^1 \cdots \omega^k$ – вектора атрибутов, которые строятся по различным методам анализа постов на стене, групп, в которых состоит пользователь, геоинформации фотографий, а также лайков, поставленных им. В зависимости от различных типов атрибутов будут

применяться различные дополнительные источники ($\omega^1 \cdots \omega^k$). Из примера можно заметить, что удалось предположить пол пользователя засчет групп, тематика которых относятся больше к мужскому полу: автомобили. Таким образом, для каждого параметра мета-профиля пользователя имеем $n+k+1$ предположения о его значениях. На основании чего возникает задача выбора одного из них. Предлагаются три подхода к решению этой задачи. В рамках данной работы не рассматриваются подходы выбора набора значений из профилей пользователей. На основании собранной информации строится распределение и выбирается значение с наибольшей частотой. В итоге для каждого параметра мета-профиля имеем четыре предположения о значениях.

IV. Подходы к решению

Предлагается три подхода к определению наиболее вероятной конфигурации набора параметров мета-профиля пользователя. Первый подход основывается на иерархии (расстановки приоритетов) различных источников информации о параметрам мета-профиля. Одним из примеров такой иерархии может быть следующий: сначала данные собираются из аккаунта пользователя, если они указаны. Если они не заданы, то выбираем одного из пользователей, находящихся в списке лучших друзей, считаем, что параметр мета-профиля целевого пользователя совпадает с ними. Если такие списки не заданы, то агрегируем данные его социального окружения (его друзей).

Второй подход строится на агрегации всех данных по какому-либо атрибуту. При слиянии вариантов атрибута у друзей и у пользователя расставляются различные коэффициенты. Полученный результат слияния по каждому атрибуту считаем окончательным ответом.

Третий подход опирается на предыдущий. Агрегация также происходит с добавлением различных коэффициентов, но в этом случае используются не все источники информации, а только некоторые. Это позволяет варьировать выборки, избегая использования источников, которые не дали никаких вариантов.

V. Алгоритм и реализация

В рамках данной работы будет рассматриваться подход, источником информации в котором выступают аккаунты пользователей в социальной сети ВКонтакте. В качестве атрибутов, для которых надо определить более вероятное значение будут выступать возраст, родной город и текущий город проживания. Определение этих трех атрибутов будет проходить при использовании следующих источников информации: личная анкета пользователя в его профиле в социальной сети; анкетные данные пользователей, являющихся друзьями исследуемого пользователя. Пусть $\hat{\omega}$ – это результат агрегации параметров мета-профилей пользователей из социального окружения исследуемого. Т.е. строится распределение по каждому параметру мета-профилей пользователей из социального окружения и выбирается максимальный по упоминанию. В конечном итоге при слияни двух оставшихся значений получается исходный параметр. Объединение происходит с использованием коэффициентов: эмпирическим путем было обнаружено, что коэффициент при результате агрегации параметров мета-профилей пользователей из социального окружения должен быть в два раза больше коэффициента при личной анкете пользователя.

VI. Итоги

Предложенный алгоритм был реализован в прототипе программного модуля. Проведено 20 тестовых запусков. В качестве параметров мета-профилей пользователей анализировались возраст, родной город и текущий город проживания. В качестве метода определения точности работы программы была выбрана кросс-валидация [4].

На выборке из 50 человек, результирующие показатели точности составили для родного города – 85,5% правильно идентифицированных аккаунтов, для текущего города проживания – 84,77% правильно идентифицированных аккаунтов, а для возраста – 91,75% правильно идентифицированных аккаунтов. Таким образам, описанный алгоритм работает со средней точностью 87.34% по всем атрибутам.

VII. Заключение

В статье предложены три подхода к решению задачи выбора наиболее вероятной конфигурации параметров мета-профиля пользователя. В качестве источника информации рассматриваются аккаунты сотрудников компании и их друзей в социальных сетях. Рассмотренные подходы закладывают основу для проведения соответствующих экспериментов, выработке новых алгоритмов. Полученные результаты будут способствовать идентификации разных аккаунтов одного пользователя в разных социальных сетях с целью агрегации большего количества информации, используемой при построении профиля уязвимостей пользователя.

Список литературы

[1] Antonyuk E.M., Varshavsky I.E., Antonyuk P.E. Adaptive systems of automatic control with prioritized channels // Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on. IEEE. 2017. P. 539–540.

[2] Bagretsov G.I., Shindarev N.A., Abramov M.V., Tulupyeva T.V. Approaches to development of models for text analysis of information in social network profiles in order to evaluate user's vulnerabilities profile //Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on. IEEE, 2017. P. 93–95.

[3] Desnitsky V.A., Kotenko I.V. Modeling and analysis of security incidents for mobile communication mesh Zigbee-based network // Soft Computing and Measurements (SCM), 2017 XX IEEE International Conference on. IEEE. 2017. P. 500–502.

[4] Langford J. Quantitatively Tight Sample Complexity Bounds. – Carnegie Mellon Thesis. 2002. 124 с.

[5] Shindarev N., Bagretsov G., Abramov M., Tulupyeva T., Suvorova A. Approach to identifying of employees profiles in websites of social networks aimed to analyze social engineering vulnerabilities // Advances in Intelligent Systems and Computing. Proceedings of the Second International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'17). Vol. 1. 2017. P.441–447.

[6] The Human Factor in IT Security: How Employees are Making Businesses Vulnerable from Within [Электронный ресурс]// Kaspersky Lab. 2017. – URL: https://www.kaspersky.com/blog/the-human-factor-in-it-security/ (дата обращения: 06.10.2017)

[7] Абрамов М.В. Автоматизация анализа социальных сетей для оценивания защищённости от социоинженерных атак // Автоматизация процессов управления. 2018. №1(51). С. 34–40.

[8] Абрамов М.В., Азаров А.А., Тулупьева Т.В., Тулупьев А.Л. Модель профиля компетенций злоумышленника в задаче анализа защищённости персонала информационных систем от социоинженерных атак // Информационно-управляющие системы. 2016. №4. С. 77-84

[9] Азаров А.А., Тулупьева Т.В., Суворова А.В., Тулупьев А.Л., Абрамов М.В., Юсупов Р.М. Социоинженерные атаки: проблемы анализа. СПб.: Наука, 2016. 352 с.

[10] Бартунов С., Коршунов А. Идентификация пользователей социальных сетей в Интернет на основе социальных связей //Труды конференции по Анализу Изображений Сетей и Текстов (АИСТ). 2012.

[11] Слёзкин Н.Е., Абрамов М.В., Тулупьева Т.В. Подход к восстановлению мета-профиля пользователя информационной системы на основании данных из социальных сетей // Сборник научных трудов Первой Всероссийской научно-практической конференции «Нечёткие системы и мягкие вычисления. Промышленные применения». (г. Ульяновск, 14-15 ноября, 2017 г.). Ульяновск, УлГТУ, 2017. С. 394–399.

[12] Социальные сети в России: исследование Mail.Ru Group: https://corp.imgsmail.ru/media/files/issledovanie-auditorij-sotcialnykh-setej.pdf