УДК 004.931
DOI: 10.26102/2310-6018/2022.38.3.002

Методы идентификации пользователей информационно-телекоммуникационной среды на основе анализа атрибутов учетных записей

А.Г. Романов✉

Академия управления МВД России, Москва, Российская Федерация
psychology.crimea@gmail.com✉

Резюме. Актуальность исследования обусловлена проблемой роста числа неустановленных лиц, совершивших преступления в сети Интернет и не только. В связи с этим, данная статья направлена на раскрытие способов установления лиц путем идентификации пользователей виртуального пространства, с целью привлечения последних к уголовной ответственности. Совершенствование информационных технологий и развитие услуг в информационно-телекоммуникационном пространстве представляют возможность анализа многочисленных данных, в том числе оставленных пользователями о себе в социальных сетях. Таким образом, ведущими методами к исследованию поставленной проблемы являются техники определения сходства буквенно-символьных объектов, созданных пользователями в атрибутах профилей социальных сетей. В настоящей статье представлен возможный алгоритм действий, направленный на деанонимизацию личности преступника. Разработка и применение методов идентификации пользователей в виртуальном пространстве позволят комплексно рассмотреть имеющуюся проблему и решить одну из основных задач, поставленных перед органами внутренних дел, связанную с раскрытием преступлений и привлечением виновных лиц к уголовной ответственности. Материалы статьи могут представлять практическую ценность для органов внутренних дел в разрезе повышения эффективности и результативности правоохранительной деятельности.

Ключевые слова: идентификация пользователя, Интернет, анализ данных, социальные сети, преступления.

Для цитирования: Романов А.Г. Методы идентификации пользователей информационно-телекоммуникационной среды на основе анализа атрибутов учетных записей. *Моделирование, оптимизация и информационные технологии.* 2022;10(3). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1203 DOI: 10.26102/2310-6018/2022.38.3.002

Methods of user identification in the information and telecommunication environment based on the analysis of account attributes

A.G. Romanov✉

Academy of Management of the Ministry of Internal Affairs of Russia, Moscow, Russian Federation
psychology.crimea@gmail.com✉

Abstract. The relevance of the study is due to the problem of the growing number of unidentified persons who have committed crimes on the Internet and beyond. In this regard, the aim of the article is to demonstrate the means for personal identification by identifying users in the virtual space in order to convict them of criminal offence. The improvement of information technologies and the development of services in the information and telecommunications space provide an opportunity to analyze numerous data, including those left by users about themselves in social networks. Thus, the leading

method to investigate the problem is the techniques to determine the similarity of alphanumeric objects created by users in the attributes of social network profiles. This article presents a possible algorithm of actions to deanonymize the identity of a criminal. The development and application of methods for identifying users in the virtual space will allow us to comprehensively consider the existing problem and accomplish one of the main tasks assigned to the internal affairs bodies and related to crime solving and charging perpetrators with a criminal offence. The materials of the article may be of practical value to the internal affairs bodies in the terms of enhancing the efficiency and effectiveness of law enforcement activities.

Keywords: user identification, Internet, data analysis, social networks, crimes.

For citation: Romanov A.G. Methods of user identification in the information and telecommunication environment based on the analysis of account attributes. *Modeling, Optimization and Information Technology*. 2022;10(3). Available from: https://moitvivt.ru/ru/journal/pdf?id=1203 DOI: 10.26102/2310-6018/2022.38.3.002 (In Russ.).

Введение

В настоящие время, сведения о действиях пользователя в сети Интернет являются ценными данными при проведении аналитических исследований в разных областях. Неизбежно, с увеличением числа пользователей, возрастает и количество данных, оставленных ими в ходе использования ресурсов сети Интернет. Набор таких данных разнообразен, в то же время структура и форма их представления зависит от источника. Полноценное исследование таких данных с учетом скорости и объема их поступления возможно при разработке новых алгоритмов и способов обработки.

Основным поставщиком данных о «жизни» пользователей в виртуальном пространстве, являются социальные сети. Взаимодействие участников между собой, а также их обращение к той или иной информации, создают предпосылки использования такой площадки для анализа. Создавая и поддерживая актуальность своего профиля, пользователь создает уникальную личность в виртуальной мире. В какой-то степени доверяя анонимности и безопасности, пользователь может опубликовать интимные сведения, касающиеся не только его, но и других пользователей. Такие сведения могут достаточно полно охарактеризовать его и окружающую его социальную среду. Отсутствие верификации размещенных сведений со стороны администрации ресурсов, дает возможность пользователю исказить данные о себе. В то же время информация о сфере интересов, области заработка и других индивидуальных чертах обычно оставляется откровенной.

Для разных целей пользователь может иметь несколько учетных записей в разных социальных сетях. В зависимости от тематической направленности или обстоятельств, при которых такие профили создаются, они могут содержать уникальные идентифицирующие признаки одного и того же пользователя. Получение и анализ такого набора данных в перспективе способен дать более развернутую характеристику пользователя. Имея в распоряжении набор сведений о лице, совершившем преступление, возможно сравнить полученные исходные данные с опорной социально-виртуальной средой. В результате, образованный уровень неопределенности в поиске преступника может значительно снизиться, повысив при этом эффективность правоохранительной деятельности. Таким образом, разработка и использование методов анализа сведений из виртуальной среды, может стать новым инструментом раскрытия преступлений.

Описание проблемы

Задача по установлению лица, совершившего преступное деяние, является основной для органов внутренних дел при раскрытии преступления. Количество

установленных лиц прямо влияет на количество привлеченных к уголовной ответственности преступников, что реализует защиту прав и интересов граждан. Однако, глобализация виртуального пространства вносит свои коррективы в правоохранительную деятельность. Применение технических и иных мероприятий не всегда дает ожидаемого результата. Анонимность действий злоумышленников в сети Интернет часто сказывается на качестве расследования уголовных дел. Таким образом, проблема деанонимизация личности преступника вызывает определенный интерес.

Накопленный научно-практический опыт предлагает ряд методов, способных помочь в решении указанной проблемы. Имеющиеся в распоряжении правоохранительных органов банки данных достаточно обширны и способствуют получению необходимых сведений. Однако наполнение и ведение таких банков происходит недостаточно динамично. Информация, содержащаяся в них, быстро устаревает и становиться не актуальной, а значит не приносит достаточного результата. С учетом информационных изменений и постоянно растущего количества пользователей, требуются новые решения, способные удовлетворить сложившуюся потребность.

В ходе раскрытия преступления устанавливается определенный набор данных о лице, его совершившим. Такими данными, являются анкетные сведения, места проживания и возможного пребывания, информация о контактах лица и его родственные связи, сведения имущественного характера, наличие судимости и иное. В зависимости от полноты и достоверности полученных сведений, составляется портрет преступника. Результативность проведения дальнейших мероприятий во многом зависит от полноты установленных сведений. Учитывая обстоятельства, связанные с эпидемиологическими обстановками по всему миру и в целом виртуализацией большинства сфер жизни, пользователи сети Интернет оставляют о себе многочисленные данные. Такие данные разнообразны и доступны для обработки во всевозможных целях. Высокая интенсивность общения и актуальное отражение социальной сферы общества на социальных платформах предоставляет возможность получать необходимые сведения для своевременного реагирования на преступные вызовы.

Становясь участником социальной сети, пользователь формирует индивидуальный профиль. Обычно, политики социальных сетей не ограничивают пользователя во вводе данных о себе. Исходя из этого предлагаемые формы профиля могут быть заполнены по усмотрению пользователя любыми сочетаниями символов и знаков, которые иногда не имеют в общественном понимании смысловой нагрузки. При этом оставленные записи могут быть понятны определенному заинтересованному кругу лиц, имеющих отношение к субкультуре или узкой профессиональной специализации. Используя такие выражения, как условные маркеры для поиска, пользователи образуют неформальные группы, в том числе и с криминальным уклоном. Преступные элементы как часть общества, достаточно часто используют социальные сети для общения. Характерной чертой такого общения является определенный сленг, представляющий собой не только известные жаргонные выражения, но и общеупотребляемые слова с заранее определенной между собеседниками смысловой нагрузкой.

Таким образом, можно сделать вывод, что предпосылками для идентификации лица как пользователя социального интерфейса являются сведения, внесенные им о себе. При создании профиля предоставляется возможность заполнения определенных форм. Заполненная пользователем форма приобретает статус атрибута и рассматривается как отдельная единица. Атрибут может иметь смысловую нагрузку, но при ее отсутствии буквенно-символьная последовательность тоже может быть использована для идентификации [1]. Стандартными примерами атрибутов могут быть антропологические данные лица, субкультурная и расовая принадлежность, локации проживания и

рождения, места обучения образования, а также иная характеризующая информация о лице. К нестандартным примерам заполнения таких форм можно отнести случаи, когда пользователь вносит сведения, понятные только ему и определенной социально-этнической группе пользователей. Также необходимо учитывать нюансы того, что многие пользователи, заполняя содержимое профилей, могут искажать данные о себе. При этом процесс идентификации таких пользователей может оказаться неточным, так как механизмы верификации обычно не предусмотрены администраторами социальных площадок. Кроме того, информация пользователями о самих себе может долгое время не обновляться и перестать быть актуальной. Такие устаревшие данные могут негативно влиять на результаты идентификации, приводя к искажениям. Помимо вышеуказанного, ограниченный доступ к аккаунту пользователя также можно рассматривать как атрибут и учитывать при исследовании.

Однако не только исследования сведений в атрибутах пользователей социальной сети могут дать положительный результат при идентификации личности. В процессе общения пользователей устанавливаются отношения, которые по степени активности и близости могут быть разными. Анализ таких отношений при помощи определенных алгоритмов и методов позволяет предпринимать попытки для деанонимизации личности пользователя. Другим предметом для исследования процессов идентификации лица могут служить непосредственные действия пользователя на платформе социальной сети. Имея в распоряжении необходимые данные о действиях пользователя, также возможно отождествить личность.

В данной работе рассматривается алгоритм идентификации пользователя по атрибутам аккаунта в социальной сети. Ожидаемым результатом является повышение эффективности и результативности правоохранительной деятельности.

Степень разработанности проблемы

Проблемой идентификации пользователей на основе анализа социальных сетей посвящено достаточное количество научных исследований. Каждое из них освещает отдельные темы общей проблематики. Так, в работе «Идентификация пользователей социальных сетей в Интернете на основе социальных связей» [2] (авторы С. Бартунов и А. Коршунов) предложена оригинальная модель «JLA» для идентификации пользователей. Основанная на модели условных случайных полей и использующая атрибуты пользовательских профилей, а также социальные связи. Данный подход может быть результативен в случаях незначительности информации, полученной из пользовательских профилей, в том числе, когда сведения недоступны или скрыты из соображений приватности. Механизм работы данной модели заключается в способности сопоставлять профили пользователей, которые трудно сравнить, используя только информацию из атрибутов.

Исследование, проведенное М.В. Лапенок и О.М. Патрушевой, посвященное теме «Идентификации пользователя в различных социальных сетях по средствам анализа социальных связей пользователя и атрибутов профиля» [3], раскрывает вопросы противодействия использованию социальных сетей в качестве средства подготовки и совершения различных видов преступной деятельности. Приведенные результаты работы алгоритма идентификации пользователя социальных сетей на основе анализа текстовых атрибутов и его социальных связей могут быть включены правоохранительными органами в состав мероприятий для борьбы с незаконной деятельностью.

Авторским коллективом А. Коршуновым, И. Белобородовым, Н. Бузун и др. рассмотрена тема «Анализа социальных сетей: методы и приложения» [4]. В работе

описаны основные компоненты стека технологий для анализа пользовательских данных из социальных сетей. Особое внимание уделяется задачам, методам и приложениям анализа сетевых и текстовых данных, таких как определение демографических атрибутов пользователей, поиск описаний событий в корпусах сообщений, идентификация пользователей различных сетей, поиск сообществ пользователей и измерение информационного влияния между пользователями. Также в работе рассмотрены подходы к получению исходных данных для анализа и сбор реальных данных путем обращения к веб-интерфейсам социальных сервисов. Для каждого из разработанных инструментов описывается его функциональность, варианты использования, основные шаги используемых алгоритмов и результаты экспериментальных исследований.

В научном сообществе Китая вопрос идентификации пользователя в социальных сетях также актуален. В статье «A Survey of Across Social Networks User Identification» [5] группой авторов Ling Xing, Kaikai Deng, Honghai Wu и др. рассмотрены вопросы идентификации пользователя по нескольким учетным записям из разных социальных сетей. Согласно проведенным исследованиям, авторы приходят к выводу, что идентификация пользователя социальной сети может помочь компании при предоставлении услуг определенному лицу. Систематически рассмотрены модели и методы идентификации, современные методы сравнения, а также их применение на практике и проблемные вопросы.

Применительно к правоохранительной деятельности использование указанных алгоритмов и методов позволит повысить ее результативность путем расширения функциональных возможностей

Описание алгоритма идентификации

В целях совершенствования деятельности органов внутренних дел представляется целесообразным разработать алгоритм действий и методов, направленных на сопоставление заданных к поиску сведений со сведениями в социальной сети. Результат работы алгоритма предполагает определение круга пользователей, удовлетворяющих запросу. Формирование запроса происходит путем составления криминалистического портрета преступника из полученных ранее сведений в ходе осуществления оперативно-розыскной деятельности. Такой портрет можно представить как универсальный шаблон с формами для заполнения. Внесенные в формы сведения образуют кортеж. Таким образом, каждый элемент кортежа содержит характеризующую единицу и рассматривается как атрибут. В свою очередь совокупность атрибутов образует криминалистический портрет. Увеличение количества атрибутов снижает уровень неопределенности об устанавливаемом пользователе, с другой стороны, информация об установленном пользователе пополняет базу знаний. В процессе поиска происходит сравнение исходных атрибутов с атрибутами пользователей целевой социальной сети. По итогам сравнения выделяется группа профилей, сформированная согласно заданному условием уровню порогового значения, рассчитанному по сумме идентичных атрибутов. При определении идентичности атрибутов предполагается использование критериев сходства.

Согласно теории графов, социальную сеть можно представить как упорядоченное множество: $G\{V,E\}$, где V – множество профилей пользователей, а E – множество отношений между пользователями. Каждая вершина v_i имеет набор других атрибутов. К этим атрибутам относится: топологическая, т. е. теоретико-графовая расположенность вершины в сети, и ее качественные характеристики, т. е. пол, возраст, город проживания и т. д.

В ходе проведения оперативно-розыскных мероприятий обычно удается установить о лице ряд сведений, представляющих интерес. В данном случае это ряд заданных для поиска атрибутов профиля в исследуемой социальной сети. Множество таких атрибутов можно преобразовать в многомерный вектор, который предоставляет возможность более точно идентифицировать пользователя, образовав выборку аккаунтов, отвечающих заданным требованиям. Таким образом, если массив сведений о пользователе будет содержать n атрибутов, то многомерный вектор возможно определить как $F_i^x = (a_1^x, a_2^x, \dots, a_n^x)$, где a_n^x обозначает n-й атрибут аккаунта.

Для того чтобы осуществить процесс идентификации, необходимо сравнить содержимое атрибутов созданного криминалистического профиля с атрибутами исследуемого профиля в целевой социальной сети. Как отмечалось выше, это может быть любая буквенно-символьная последовательность. Когда пользователь, оставляя о себе сведения, формирует аккаунт социальной сети, он употребляет слова и выражения, которые ему привычны. В совокупности внесенные сведения образуют его образ в виртуальной среде, который презентуется другими участникам в целях общения. Свобода Интернета в какой-то степени раскрепостила пользователей и нередко обращается внимание на то, что люди ведут себя более раскованно на информационных просторах. В этой связи, определенным лингвистическим индикатором, характеризующим пользователя, может служить частота использования одного слова или выражения. При этом частота употребления определенного слова рядом пользователей может служить определенным маркером распространения в обществе новых идей или тенденций. Указанные единицы языка могут иметь общую тематическую окраску и ярко выражать принадлежность к среде, что может быть использовано в процессе идентификации.

Таким образом, содержанием одного из атрибутов вектора F_i^x может являться специальный термин. Вычислить чистоту использования данного термина можно при использовании статистической меры TF-IDF (term frequency – inverse document frequency – частота терминов – обратная частота документа), используемой для количественной оценки важности или релевантности строковых представлений [6].

Так, первоначально вычисляется частота использования каждого слова в исследуемом профиле пользователя социальной сети

$$TF = \frac{n}{N}, \quad (1)$$

где n – количество употребления определенного слова, а N – общее количество слов. Далее вычисляется обратная частота для каждого слова в профиле пользователя

$$IDF = \log\left(\frac{D}{P+1}\right), \quad (2)$$

где D – общее количество профилей пользователей социальной сети, P –количество профилей, содержащих искомое слово и добавляется "1", чтобы избежать случаев, когда знаменатель равен 0. После этого производится вычисления для каждого слова в профиле

$$TF\text{-}IDF = TF \times IDF = \frac{n}{N} \times \log\left(\frac{D}{P+1}\right). \quad (3)$$

Произведя выборку ключевых слов в каждом из сравниваемых профилей, полученные данные образуют векторы частоты употребления слов A_i и B_i для дальнейшего вычисления сходства. Для этого предлагается использование косинусного сходства как меры подобия двух векторов [7]. При этом необходимо учитывать, что при вычислении мера изменяется в диапазоне от 0 до 1, поскольку частота использования не

может быть отрицательной. Для получения результата косинусного сходства используем векторы A_i и B_i

$$cos(0) = \frac{\sum_{i=1}^{k}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{k}(A_i)^2} \times \sqrt{\sum_{i=1}^{k}(B_i)^2}}. \quad (4)$$

Точность процесса идентификации зависит от количества совпадающих значений в атрибутах профилей социальной сети. Вес таких значений может быть разным, но обычно пользователи оставляют достаточно информации о себе. Такие данные можно представить как множество атрибутов. Таким образом, для сравнения сведений между заданными профилями возможно построить вектор сходства. Вектор можно определить как $V\ (F_A, F_B) = (v_1^{AB}, v_2^{AB}, \ldots, v_n^{AB})$, где v_i^{AB} обозначает сходство n-го атрибута между исследуемым профилем целевой социальной сети и заданным криминалистическим профилем устанавливаемого лица. Таким образом, идентификация пользователя будет происходить в результате вычислений сходства между двумя профилями, где один кортеж сведений в атрибутах сравнивается с другим.

Методы реализации алгоритма идентификации

Реализация отождествления личности пользователя возможна при использовании функций подобия. Известны различные способы вычисления сходства, такие как коэффициент Жаккара, коэффициент Кульчинского, коэффициент Отиаи, коэффициент Шимкевича, коэффициент Браун-Бланке, коэффициент Серенсена-Дайса, расстояние Левенштейна и расстояние Джаро-Винклера [8]. Обратимся к некоторым из них для вычисления значения сходства между последовательностями в атрибутах профилей.

Коэффициент Серенсена-Дайса является статистической мерой и используется для оценки сходства двух выборок. Первоначально формула предназначалась для применения к дискретным данным, где учитывалось два набора X и Y, он определялся как

$$DSC = \frac{2|X \cap Y|}{|X|+|Y|}, \quad (5)$$

где $|X|$ $и$ $|Y|$ являются количествами элементов в каждом наборе. Коэффициент равен удвоенному количеству элементов, общих для обоих наборов, деленному на сумму количества элементов в каждом наборе.

Хотя обоснование использования данного коэффициента в первую очередь эмпирическое, теоретически его можно обосновать как пересечение двух нечетких множеств. Теория нечетких множеств позволяет равномерно оценивать принадлежность элементов к множеству. Описывая при помощи функции принадлежности, имеющие значения в действительном единичном интервале [0,1]. Образованные двухвалентные наборы называются четкими наборами. Теория нечетких множеств может использоваться в широком диапазоне областей, в которых информация является неполной или неточной. Возможно использовать данный метод для решения задачи идентификации. При этом сведения в атрибутах сравниваемых аккаунтов можно представить в виде многозначных строк. Так, коэффициент будет равен удвоенному количеству элементов, общему для обоих наборов, деленному на сумму количества элементов в каждом наборе.

Представим информацию, содержащуюся в двух сравниваемых атрибутах профилей как элементы строк v_i и v_j. Удвоенное количество пересечений элементов строк, деленное на сумму элементов в каждой строке v_i и v_j, дает две строки

коэффициентов, с помощью которых снова вычисляются строки коэффициентов определяя тем самым сходство двух атрибутов

$$(F_A, F_B) = 2\frac{|v_i \cap v_j|}{|v_i| + |v_j|}. \quad (6)$$

Применение данного метода может быть следующим. Предположим, что необходимо установить пользователей социальной сети, распространяющих наркотики. При создании криминалистического профиля устанавливаемого лица в один из атрибутов поиска можно внести словосочетание «Распространяю наркотики всем». Таким образом в целевой социальной сети будет определяться сходство сведений, оставленных пользователями в атрибутах своих профилей. Можно предположить, что у одного из пользователей социальной сети атрибут имеет строку следующего содержания «Всем привет». Результаты использования метода определят сходство равное 2/5 = 0,4. Так как информация о пресечении в строках атрибутов равна 2, общие количество слов в двух строках равно 5.

Такие вычисления будут справедливы и для разбора слова. При этом вычисления происходят с использованием биграмм, разделяя слово на пары букв и сравнивая их между заданными словами следующим образом:

$$(F_A, F_B) = \frac{2v_t}{v_x + v_y}, \quad (7)$$

где v_t – количество символьных биграмм, найденных в обеих строках, v_x – количество биграмм в первой строке, а v_y – количество биграмм во второй строке. Например, вычислим сходство между словами «хакер» и «покер». В каждом слове мы найдем набор биграмм {ха, ак, ке, ер} и {по, ок, ке, ер}. Каждый набор состоит из 4 элементов, а пересечение этих наборов имеет два элемента: {ке} и {ер}. В результате коэффициент сходства двух слов будет равно 4/8 = 0,5.

Другим методом для определения сходства является расстояние Левенштейна, которое измеряет по модулю разность между двумя последовательностями символов [9]. Расстояние определяется как минимальное количество одно символьных операций, необходимых для превращения одной последовательности символов в другую. В теории информации и компьютерной лингвистики применимо название редакционного расстояния, поскольку в процессе измерений вычисляется минимальное количество действий по редактированию одной строки для получения другой. Применимы следующие операции: вставка или удаление символа, замена одного символа другим.

$$F_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0, \\ \min \begin{cases} F_{a,b}\,(i-1,j) + 1 \\ F_{a,b}(\,i,j-1) + 1 \\ F_{a,b}(i-1,j-1) + \ 1_{(a_i \neq b_j)} \end{cases} \end{cases} \quad (8)$$

Расстояние между строками a и b (длины $|a|$ и $|b|$ соответственно) задается как $F_{a,b}(|a|,|b|)$, где, $1_{(a_i \neq b_j)}$ – индикаторная функция, равная 0, когда $a_i = b_j$, и равная 1 в противном случаи, и $F_{a,b}(i,j)$ – расстояние между первыми i символами строки a и первыми j символами строки b. При этом индексы i и j отсчитываются от единицы. При этом первый элемент в min соответствует операции удаления символов от строки a, символов строки b. Второй элемент соответствует вставке символов, а третий – для совпадения или несоответствия, в зависимости от того, совпадают ли соответствующие символы. Использование данного метода для определения сходства возможно при вычислении расстояния от каждого вхождения элемента заданных сведений в атрибутах

созданного криминалистического профиля к элементам атрибутов профилей опорной социальной сети. В результате значение 0 указывает на то, что исходные заданные сведения абсолютно аналогичны сведениям исследуемого профиля опорной социальной сети, а для примера значение 1 также укажет на аналогичные сведения, но с отличием на 1 символ.

На основании данного метода, идентификацию пользователя возможно осуществить несколькими способами: рассчитывая пословное расстояние, рассчитывая взвешенное расстояние, а также производя вычисления расстояния Дамерау-Левенштейна.

При вычислении расстояния пословно, за единицу принимается не один символ, а одно слово. Таким образом, возможно измерить сходство между двумя словосочетаниями в двух атрибутах. Так, имя два выражения «Я использую программы» и «Я использую хакерские программы». Превратить первое выражение во второе, возможно за 1 операцию – вставка слова «хакерские». В случае вычислений посимвольно, пришлось бы совершить 11 операций – вставка, учитывая количество букв в слове и пробелы.

При расчете взвешенного расстояния необходимо учитывать, что, когда определенное количество редакционных операций, преобразовывает одно выражение в другое, такие операции могут соответствовать различным выравниваниям. Чтобы сравнить количественные результаты разных вариантов выравнивания, каждая операция соотносится с определенной системой оценок. Данная система может быть элементарной. Так, например, прибавление двух баллов за каждую пару совпадающих символов при операции замена или штраф, за отнимание трех баллов при операции несовпадения. Такие штрафы могут быть фиксированные, пропорциональные и линейные. Необходимо отметить, что системы оценок могут быть разными. Но неизменным условием остается наличие некоторой суммарной оценки, выраженной в счете или весе, которая сопоставляется с каждым выравниванием. Тем самым можно говорить о количественном показателе подобия выравниваемых последовательностей операций.

Расстояние Дамерау-Левенштейна является мерой для нахождения схожести двух строк путем реализации нечеткого поиска. В отличие от классического представления расстояния Левенштейна, помимо трех основных операций редактирования используется транспозиция. Благодаря перестановке двух соседних символов при определении сходства удается повысить эффективность сравнения, так как при внесении пользователями данных самыми распространенными искажениями являются опечатки и орфографические ошибки.

Еще одним методом определения сходства является расстояние Джаро-Винклера, измеряющего расстояние редактирования между двумя последовательностями на основе расстояния Джаро, определяющего расстояние между двумя заданными строками [10]. Расстояние Джаро-Винклера, используя коэффициент масштабирования, предлагает более расширенные критерии оценки для совпадающих строк. Две строки будут более схожи при меньшем расстоянии Джаро-Винклера. Оценка нормализована так, что ноль указывает на полное совпадение, а один – на отсутствие сходства.

Так, для определения сходства $F_{a,b}$, представим сведения в атрибуте созданного криминалистического профиля как s_1, а сведения в атрибуте исследуемого профиля целевой социальной сети как s_2, тогда:

$$F_{a,b} = \begin{cases} 0 \quad \text{когда } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), \end{cases} \tag{9}$$

где $|s_1|-$ длина строки атрибута криминалистического профиля, $|s_2|$ – длина строки атрибута профиля опорной сети, $m-$ число совпадающих символов, а t – половина числа транспозиций (перестановки).

Два символа из строк s_1 и s_2 соответственно, считаются схожими, только если они одинаковы, и их позиция относительно друг друга находятся не дальше, чем на d – максимальное расстояние для поиска:

$$d = \left\lfloor \frac{max(|s_1|,|s_2|)}{2} \right\rfloor - 1 \quad (10)$$

Каждый символ строки s_1 сравнивается со всеми соответствующими ему символами в сроке s_2 на допустимое расстояние d. Количество совпадающих, но отличных порядковыми номерами символов делится на два, определяя при этом число перестановок.

Для примера представим две строки ddwrit.exe и jwwdti.exe. (Рисунок 1) Выполняя расчет расстояния, получим максимальное расстояние для поиска перестановок, которое составляет 10/2-1=4 число совпадающих символов, равное 5, число несовпадающих символов, равное 7 и половину числа транспозиций, равное 3. Расстояние Джаро для двух строк равняется 0.46 (Рисунок 2).

	d	d	w	r	i	t	.	e	x	e
j	0	0	0	0	0	0	0	0	0	0
w	0	0	**1**	0	0	0	0	0	0	0
w	0	0	1	0	0	0	0	0	0	0
d	**1**	**1**	0	0	0	0	0	0	0	0
t	0	0	0	0	0	**1**	0	0	0	0
i	0	0	0	0	**1**	0	0	0	0	0
.	0	0	0	0	0	0	1	0	0	0
e	0	0	0	0	0	0	0	1	0	**1**
x	0	0	0	0	0	0	0	0	1	0
e	0	0	0	0	0	0	0	**1**	0	1

Рисунок 1 – Матрица расчета расстояния
Figure 1 – Distance calculation matrix

Как указывалось выше, при вычислении расстояния Джаро-Винклера используется коэффициент масштабирования – p, что дает благоприятные рейтинги строкам, которые совпадают друг с другом от начала до определенной длины ℓ, которая называется префиксом.

Так, даны две строки $\mathcal{S}_1$ и $\mathcal{S}_2$, их расстояние d_w следующее

$$d_w = d_j + (\ell p(1 - d_j)), \quad (11)$$

где d_j- расстояние Джаро для строк $\mathcal{S}_1$ и $\mathcal{S}_2$, $\ell-$ длина общего префикса от начала строки до максимума четырех символов, p – постоянный коэффициент масштабирования, использующийся для того, чтобы скорректировать оценку в сторону повышения для выявления наличия общих префиксов, при этом не должен превышать 0,25 в противном случае расстояние может стать больше, чем 1. Стандартное значение этой константы равно 0.1. В конкретном примере длина общего префикса равна 0, поэтому применение данного метода не целесообразно.

D	$м$	S_1	S_2	tm	t	ℓ	d_j	d_w
4	5	10	10	7	3	0	0.46	0.46

Рисунок 2 –Таблица результатов
Figure 2 – Result table

Заключение

На основании вышеизложенного можно сделать вывод, что использование методов идентификации пользователей социальных сетей на основе определения сходства значений атрибутов потенциально позволит повысить результативность и эффективность правоохранительной деятельности в аспекте решения задач по выявлению лиц, занятых в криминальной деятельности, установление которых на сегодняшний день в основном происходит в «ручном», не автоматизированном режиме. Разработка и применение способов деанонимизации пользователей виртуального пространства, имеет достаточно высокий потенциал для развития в ведомственной научной-практической среде. Помимо реализации методов идентификации на практике, основной задачей в процессе установления личности преступника будет создание и поддержание в актуальном состоянии базы знаний, в которой могут содержаться значения атрибутов пользователей ранее привлекавшихся к уголовной ответственности по определенным составам преступлений. Кроме этого, представляется целесообразным совместить данную базу знаний с уже имеющимися ведомственными массивами данных, что позволит усовершенствовать деятельность органов внутренних дел в еще большей мере.

СПИСОК ИСТОЧНИКОВ

1. Perito D., Castelluccia C., Kaafar M.A., Manils P. How unique and traceable are usernames. *International Symposium on Privacy Enhancing Technologies Symposium.* 2011;1–17.
2. Бартунов С., Коршунов А. Идентификация пользователей социальных сетей в Интернет на основе социальных связей. *Труды Института системного программирования Российской академии наук*. 2012; 14(2):1–13.
3. Лапенок М.В., Патрушева О.М. Идентификация пользователя в различных социальных сетях по средствам анализа социальных связей пользователя и атрибутов профиля. *Образовательные технологии и общество*. 2016;19(3):584–594.
4. Коршунов А., Белобородов И., Бузун Н., Аванесов В., Пастухов Р., Чихрадзе К., Козлов И., Гомзин А., Андрианов И., Сысоев А., Ипатов С., Филоненко И., Чуприна К., Турдаков Д., Кузнецов С. Анализ социальных сетей: методы и приложения. *Труды Института системного программирования Российской академии наук.* 2014;26(1):439–456.
5. Ling Xing, Kaikai Deng, Honghai Wu и др., A Survey of Across Social Networks User Identification. *IEEE Access.* 2019;7:137472–137488.
6. Karen S.J. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation.* 2004;60(5):493–502.
7. Гайдамакин Н.А. Мера сходства последовательностей одинаковой размерности. *Математические структуры и моделирование.* 2016;4(40):5–16.
8. 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;3:29–43.
9. Леонтьев В.К. О мерах сходства и расстояниях между объектами. *Журнал вычислительной математики и математической физики*. 2009;49(11):2041–2058.
10. Понизовкин Д.М. Влияние меры сходства на результативность РС. *Программные системы: теория и приложения*. 2014;5(23):55–65.

REFERENCES

1. Perito D., Castelluccia C., Kaafar M.A., Manils P. How unique and traceable are usernames. *International Symposium on Privacy Enhancing Technologies Symposium.* 2011;1–17.
2. Bartunov S., Korshunov A. Identification of users of social networks on the Internet based on social connections. *Trudy Instituta sistemnogo programmirovaniya Rossiyskoy akademii nauk = Proceedings of the Institute of System Programming of the Russian Academy of Sciences*. 2012;14(2):1–13. (In Russ.)
3. Lapenok M.V., Patrusheva O.M. User identification in various social networks by means of analyzing the user's social connections and profile attributes. *Obrazovatel'nyye tekhnologii i obshchestvo = Educational technologies and society*. 2016;19(3):584–594. (In Russ.)
4. Korshunov A., Beloborodov I., Buzun N., Avanesov V., Pastukhov R., Chikhradze K., Kozlov I., Gomzin A., Andrianov I., Sysoev A., Ipatov S., Filonenko I., Chuprina K., Turdakov D., Kuznetsov S. Analysis of social networks: methods and applications. *Trudy Instituta sistemnogo programmirovaniya Rossiyskoy akademii nauk = Proceedings of the Institute of System Programming of the Russian Academy of Sciences.* 2014;26(1):439–456. (In Russ.)
5. Ling Xing, Kaikai Deng, Honghai Wu и др. A Survey of Across Social Networks User Identification. *IEEE Access.* 2019;7:137472–137488.
6. Karen S.J. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation.* 2004;60(5):493–502.
7. Gaidamakin N.A. Measure of similarity of sequences of the same dimension. *Matematicheskiye struktury i modelirovaniye = Mathematical structures and modeling*, 2016;4(40):5–16. (In Russ.)
8. Korepanova A.A., Oliseenko V.D., Abramov M.V., Tulupyev A.L. Application of machine learning methods in the task of identifying user accounts in two social networks. *Komp'yuternyye instrumenty v obrazovanii = Computer Tools in Education.* 2019;3:29–43. (In Russ.)
9. Leontiev V.K. About similarity measures and distances between objects. *Zhurnal vychislitel'noy matematiki i matematicheskoy fiziki = Journal of Computational Mathematics and Mathematical Physics.* 2009;49(11):2041–2058. (In Russ.)
10. Ponizovkin D.M. The influence of similarity measures on the effectiveness of RS. *Programmnyye sistemy: teoriya i prilozheniya = Software systems: theory and applications*. 2014;5(23):55–65. (In Russ.)

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Романов Александр Георгиевич, адъюнкт, кафедра информационных технологий, факультет подготовки научных и научно-педагогических кадров, Академия управления МВД России, Москва, Российская Федерация. *e-mail:* psychology.crimea@gmail.com

Alexander G. Romanov, Postgraduate Student, Department of Information Technologies, Faculty of Scientific and Scientific-Pedagogical Personnel Training, Academy of Management of the Ministry of Internal Affairs of Russia, Moscow, Russian Federation.

Статья поступила в редакцию 15.06.2022; одобрена после рецензирования 29.06.2022; принята к публикации 15.07.2022.

The article was submitted 15.06.2022; approved after reviewing 29.06.2022; accepted for publication 15.07.2022.