Анализ персональной информации из социальных сетей для решения задач криминалистики

Е.А. Гамбарова[1], В.С. Бакаев[1], Н.В. Олиндер[1], А.В. Благов[1], М.Е. Наумов[1]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

Аннотация. В статье обсуждается необходимость использования социальных сетей в познавательной деятельности участников уголовного процесса и предлагается использовать информацию, полученную из социальных сетей, при расследовании преступлений. Дается сравнение двух подходов: экспертного и автоматизированного. Авторы предлагают инструменты для сбора и анализа персональных данных из социальных сетей.

1. Введение

Важным условием повышения эффективности противодействия современной преступности является постоянное совершенствование теоретических и практических знаний следователя, применение современных технологий в расследовании, поиска новых путей сбора информации.

Информация занимает центральное место в познавательной деятельности следователя, поэтому поиск путей более быстрого и полного получения информации является важным направлением в криминалистике. Большие возможности для работы с информацией предоставляет сеть Интернет, в частности, социальные медиа (социальные сети).

В настоящее время необходимо отметить изменение подхода граждан к способам и формам общения, обороту информации и т.п., отчасти это обусловлено развитием глобальной сети Интернет, развития виртуальных отношений. Все больший оборот набирает общение посредством социальных сетей и мессенджеров [1]. Рост популярности социальных медиа (социальных сетей, мессенджеров и пр.) [2], широкое распространение «виртуальных» баз данных, он-лайн банкинга, облачного хранения и других инструментов, используемых с целью более удобного и быстрого построения коммуникаций и получения (предложения) услуг приводит не только к необходимости нормативного регулирования данных отношений, но и обуславливает создание новых подходов к работе с виртуальным пространством в криминалистике.

При планировании отдельных следственных действий перед следователем стоит задача подобрать наиболее эффективные способы достижения поставленной цели. Как правило, одной из задач планирования следственных действий является сбор информации, которая в дальнейшем поможет следователю выбрать тактику проведения того или иного следственного действия. В связи с тем, что далеко не всегда у следователя имеется большое количество времени для поиска информации об интересующем событии или личности, необходимо подобрать способы, которые помогут сократить время получения информации. Представляется

перспективным использование информации из интернета, в том числе проводя мониторинг социальных сетей, при планировании следственного действия, в частности, допроса.

2. Методы сбора и обработки данных социальных сетей

В криминалистике при организации расследования бывает необходимо оперативно получить достоверную информацию о человеке или группе людей. Оперативность, достоверность и своевременность при этом являются ключевыми факторами. Поэтому представляется целесообразным разработать технологию наиболее эффективного получения и обработки информации.

Задача сбора необходимой информации из социальных сетей может быть поделена на сбор данных, их фильтрацию, обработку и последующий анализ.

Авторами исследования была поставлена задача сбора данных социальных сетей экспертно, а также при использовании разработанных программных средств. В первом случае определялась группа людей экспертов, которые системно без применения дополнительных автоматизированных сервисов искали нужную персональную информацию. Во втором случае было использован следующий подход.

Исходя из поставленной задачи, разработанный программный комплекс реализует следующий функционал:

1. Анализ всех профилей целевых социальных сетей (ВКонтакте, Twitter, Instagram, LinkedIn) с целью сохранения открытой информации в базу данных.
2. Объединение профилей, принадлежащих одному человеку, в группы.
3. Выдвижение предположений об уровне дохода пользователя.

Для реализации программного продукта был использован следующий стек технологий: Scala, Python, PostgreSQL, Apache Storm, CatBoost. Данный выбор обусловлен требованием к горизонтальному масштабированию системы.

CatBoost используется для построения математическое модели, определяющий уровень дохода человека по таким параметрам, как: пол, возраст, образование, сфера деятельности, должность, город, семейное положение.

Для поиска аккаунтов человека в других социальных сетях мы используем двухслойный персептрон, сравнивая профили по имени, никнейму, электронной почте и т.п.

Группирование данных основывается на анализе общих черт [3]. При этом применяется следующая процедура:

Построение полного многодольного графа, в котором хранится информация о профилях социальных сетей и потенциал, характеризующий вероятность их принадлежности одному человеку;

В вершинах графа содержится информация о профилях, которая используется при их сравнении.

Для сравнения двух профилей используется многослойная нейронная сеть. На входной слой сети подаётся вектор размерности 12, содержащий следующие данные:

- Name ↔ Name'
- max(Name → Username', Name' → Username)
- max(Name → E-mail', Name' → E-mail)
- max(Name → Skype', Name' → Skype)
- Username ↔ Username'
- max(Username → E-mail', Username' → E-mail)
- Username ↔ Skype'
- max(Skype → Username', Skype' → Username)
- max(Skype → E-mail', Skype' → E-mail)
- E-mail ↔ E-mail'
- Phone ↔ Phone'
- Website ↔ Website'
-

Далее определяется Полнота вхождения a в b:

$$a \to b = 1 - \frac{d + r + s}{len(a)} \in [0, 1],$$

где d - количество операций удаления для преобразования a в b; r - количество операций замены для преобразования a в b; s - количество операций транспозиции для преобразования a в b; *len(x)* - функция вычисления длины аргумента.

Сравнение a и b:

$$\forall i \in [1, len(a)], j \in [1, len(b)] d[i, j] = 1 - \frac{dist(a[i], b[j])}{len(b[j])} \in [0, 1],$$

$$a \leftrightarrow b = \frac{\sum_1^{len(a)} d[i, fit(i)]}{min(len(a), len(b))} \in [0, 1],$$

где *dist(a, b)* - функция, вычисляющая расстояние Дамерау-Левенштейна [4] для строк a и b; *fit(i)* - функция, возвращающая индекс слова строки b, поставленного в соответствие слову *a[i]*.

Операция сравнения не учитывает порядок слов. Все слова исходных строк попарно сравниваются, а затем, при помощи алгоритма Куна-Манкреса [5], каждому слову строки a ставится в соответствие слово строки b так, чтобы сумма схожести по всем парам слов была максимальной. Также не учитываются знаки препинания и прочие символы (за исключением букв и цифр).

Обучающая и контрольная выборки собраны на основе первичных данных. Размер обучающей выборки ~106 пар.

Далее в сгенерированном графе для каждой пары долей выполняется следующая последовательность действий:

- рёбра сортируются в порядке убывания весов;
- удаляются рёбра, вес которых меньше порогового значения или одна из инцидентных вершин уже связана с какой-либо вершиной противоположной доли.

В результате этих преобразований получается граф, в котором каждая компонента связности представляет собой группу аккаунтов из разных социальных сетей, которые принадлежат одному человеку.

В связи с тем, что человек может одновременно принадлежать нескольким сообщества, а также если одна и та же группа была сформирована в нескольких сообществах, то можно полагать, что аккаунты этой группы действительно принадлежат одному пользователю.

Для парсинга контактной информации применяются регулярные выражения и встроенный в Scala механизм работы с КС-грамматиками.

3. Результаты и обсуждения

При использовании способа экспертного сбора и обработки данных социальных сетей быди получены следующие результаты. Был проведен следующий эксперимент. Экспертной группе (115 человек), предложили искать информацию об определённых людях (3 человека) по заданным параметрам: место жительства лица, место учёбы лица, вхождение в состав учредителей, наличие имущества, наличие задолженностей и штрафов, участие в судебных процессах, путешествия и деловые поездки, досуг, состав семьи, близкие друзья.

Важным условием было, чтобы участники эксперимента не использовали специальных технических средств и искали информацию только в открытых источниках.

В результате эксперимента выяснилось, что с лёгкостью можно найти информацию о городе проживания (67% участников нашли), хотя конкретный адрес удалось найти только 5,7% участников. Место работы и учёбы смогли обнаружить 79% участников. Вхождение в состав учредителей, акционеров, наличие статуса индивидуального предпринимателя и т.д. нашли 22% участников эксперимента, наличие информации об имуществе смогли найти только 10% участников. Задолженности, штрафы и кредиты нашли менее 2% участников, участие в судебных процессах (в качестве стороны) нашли менее 1%, путешествия и деловые поездки смогли найти более половины участников эксперимента - 50,6%, информацию о родителях смогли найти 29,5% участников эксперимента, о супругах более половины участников 52,5%; о братья и сестрах 19% , о друзьях в среднем 20,9 %.

Выполнение работ по поиску и обработке информации (в том числе определении её достоверности заняло также в среднем два с половиной часа).

Вторым способом с использованием разработанного программного средства было получено следующее. Созданная система построения портретов пользователей социальных сетей способна собирать следующие данные:

- идентификаторы профилей пользователя в других социальных сетях (в том числе, если он их не указывал явно на своей странице)
- прочая контактная информация (номера телефонов, адреса электронной почты, логины Skype)
- ФИО пользователя и никнеймы
- дата рождения
- город проживания
- родственники (родители, дети, братья, сестры)
- образование (университет, школа)
- место работы и должность
- уровень дохода (с использованием статистики сервисов HeadHunter и Яндекс.Работа)

В качестве эксперимента были проанализированы участники сообщества «Большая деревня» (https://vk.com/bigvill). За 197 секунд обработала данные 48.525 профилей ВКонтакте (Instagram – 8734, Twitter – 4367, LinkedIn – 1455).

В результате можно увидеть, что с помощью программного средства можно гораздо более оперативно собрать и обработать большую информацию, при этом, конечно же, более детальный анализ можно проводить экспертно. Разработанный программный продукт может использоваться в криминалистике для оперативного предварительного анализа персональных данных, в том числе для проверки их достоверности (по различным параметрам, к примеру, соответствии указанных дат).

В целом можно сказать, что социальные сети могут рассматриваться для решения задач криминалистики и служить объектом исследования. Открытость и как следствие доступность данных с одной стороны, это является негативным фактором, так как снижен уровень защиты персональных данных (хотя эти данные размещают сами субъекты). С другой стороны, такая «открытость» может помочь в работе следственных органов при расследовании преступлений. Например, при сборе информации о возможных участниках преступных группировок, при подготовке к отдельным следственным действиям (например, допрос, очная ставка) или в целом, при планировании расследования отдельных видов преступлений.

4. Выводы

Результатом работы является исследование сбора и обработки персональных данных пользователей социальных сетей для решения задач криминалистики.

Очевидно, что сравнение производительности и качества информации, полученной человеком и машиной, дает предсказуемый результат. Однако использование правоохранительными органами подобного программного средства может в значительной мере сократить время поиска основной информации и отбросить из выборки людей, не удовлетворяющих заданным критериям. А далее дополнительная информация может быть собрана экспертно.

5. Благодарности

Работа выполнена при частичной финансовой поддержке Министерства образования и науки РФ в рамках реализации Программы повышения конкурентоспособности Самарского университета среди ведущих научно-образовательных центров мира на период 2013-2020-х годов.

6. Литература

[1] Dupuis, M. “I Got the Job!”: An exploratory study examining the psychological factors related to status updates on facebook / M. Dupuis, S. Khadeer, J. Huang // Computers in Human Behavior. – 2017. – Vol. 73. – P. 132-140.
[2] Олиндер, Н.В. О результатах эксперимента «поиск и восприятие информации о личности в сети интернет и ее использование при расследовании преступлений / Н.В. Олиндер, Е.А. Гамбарова // Эксперт-криминалист. – 2017. – №4. – С. 29-31.
[3] Bakaev, V.A. The analysis of profiles on social networks / V.A. Bakaev, A.V. Blagov // CEUR Workshop Proceedings. – 2017. – Vol-1903. – P. 88-91.
[4] Smetanin, N. Fuzzy search in the text and the dictionary / N. Smetanin. – [Electronic resource]. – Access mode: https://habrahabr.ru/post/114997 (in Russian).
[5] Hungarian algorithm for solving the assignment problem. [Electronic resource]. – Access mode: http://e-maxx.ru/algo/assignment_hungary (in Russian).

Analysis of the personal information from social networks to solve the problems of criminology

E.A. Gambarova[1], V.S. Bakaev[1], N.V. Olinder[1], A.V. Blagov[1], M.E. Naumov[1]

[1]Samara National Research University, Moskovskoe shosse 34A, Samara, Russia, 443086

Abstract. The article discusses the need to use social networks in the cognitive activities of participants in the criminal process, and suggests that it is possible to use information obtained from social networks in the investigation of crimes. Two approaches are compared: expert and automated. The authors offer tools for data collection and analyzing personal data from social networks.

Keywords: data mining, social networking, social networks, security and privacy, crime investigation, cyber-investigation, personality, digital evidence.