Гревцов Д.В.

*доцент кафедры защиты информации Военно-инженерного института радиоэлектроники и связи МО РК, города Алматы*

**ЦИФРОВЫЕ ОТПЕЧАТКИ НА СТРАЖЕ ПЕРСОНАЛЬНЫХ ДАННЫХ**

Всякий раз, обращаясь в банк, поликлинику, страховую компанию и множество других организаций, мы заполняем анкеты. Много разных анкет. И везде собственноручно или с помощью оператора мы вносим в эти анкеты персональные данные. Это могут быть заявления на кредит, согласие на обработку персональных данных, типовые формы различных заявлений и договоров и т. п. Что происходит с ними потом?

Заявления и анкеты с персональными данными граждан, предоставляемые по требованию разных организаций, в некоторых случаях так и остаются в бумажном виде без перевода в цифровой. Однако стандартной практикой современных организаций является заполнение форм в электронном виде (если предоставленные данные сразу вносятся в информационную систему организации), когда клиент подписывает готовую форму, или перевод заполненных документов в электронный вариант — довольно часто подписанная анкета сразу фотографируется или сканируется. Также персональные данные присутствуют в различных договорах, заключаемых между гражданами и организациями. Кроме того, к заявлениям граждан нередко прилагаются сканы документов — паспортов, водительских удостоверений и т. п. Утечка собранных персональных данных, разумеется, недопустима ни в каком виде, однако такие случаи не редкость.

С помощью поисковых серверов можно легко найти в свободном доступе огромное количество документов: так, недавно в Сеть утекли заявления и документы Управления социальной защиты населения по городскому округу Балашиха, другим поисковым запросом можно было найти обращения граждан в Департамент транспорта Москвы, часто с прикрепленными сканами документов (паспорта, водительские удостоверения и т. п.). Известны и случаи утери заполненных анкет в бумажной форме (когда, например, «незначительная часть документов была по ошибке выброшена вместе с устаревшими рекламными материалами» у офиса Сбербанка).

Риск намеренной или случайной утечки персональных данных, равно как и конфиденциальной корпоративной информации, всегда присутствует и должен быть учтен в корпоративной политике безопасности. Заявления. Анкеты. Договоры. Что объединяет эти виды документов?

Все они создаются на базе унифицированных форм. Для выявления, инспекции и предотвращения утечки такого рода данных можно использовать различные технологии контентной фильтрации — искать в документах определенные слова или их сочетания, выявлять некоторые персональные данные с помощью шаблонов регулярных выражений, а также использовать технологию цифровых отпечатков (фингерпринтов).

Типовые документы, описанные в данной статье, обладают уникальными словесными шаблонами. Пустая форма заявления или договора отличается от заполненной только наличием дополнительных сведений, содержит больше текстовых или графических данных по сравнению с образцом-шаблоном. Каждый раз при заполнении формы используется тот же исходный набор слов и данных, к которому затем добавляются собственные слова и изображения (например, подпись или печать). Технология цифровых отпечатков позволяет выявить уникальный словесный шаблон в отправляемом по каналам сетевых коммуникаций, сохраняемом на внешний накопитель или отправленном на печать документе, сравнивая документ с имеющимся шаблоном, а точнее, с цифровым отпечатком базового документа — формы.

 По сравнению с методами анализа по ключевым словам или шаблонам регулярных выражений метод цифровых отпечатков является наиболее эффективным и позволяет обеспечить наименьшее число ложных срабатываний, поскольку контролируемый образец всегда входит в исходящий документ, и, как следствие, идентифицируется соответствие отпечатку. В основе метода цифровых отпечатков лежит сопоставление документов или файлов с так называемыми цифровыми отпечатками — наборами буквенно-цифровых строк (хэшей), при помощи которых можно идентифицировать данные, хранящиеся в документе или файле.

Логика проверки соответствия цифровых отпечатков документа и шаблона Каждый шаблон (форма) содержит описания полей, некий набор текста и пустые поля для их дальнейшего заполнения. Преобразование такого шаблона в отпечаток документа и сохранение его уникального хэш-значения для последующей классификации в базе данных цифровых отпечатков позволяет получить тип конфиденциальных данных, который затем DLP-система может сопоставить с передаваемым документом. Что важно, сам исходный документ в базе данных отпечатков не сохраняется и воссоздать его из цифрового отпечатка невозможно. При сопоставлении заполненного документа с отпечатком шаблона документа DLP-система определяет, какова степень соответствия отпечатков передаваемого документа и шаблона, и применяет заданную политику защиты от утечки данных. Такая политика может, например, запрещать пользователям отправлять исходящие сообщения с персональными данными в защищаемых формах на внешние (не корпоративные) адреса электронной почты. Метод цифровых отпечатков можно успешно применять к большинству текстовых форм, используемых в организациях, например, бланкам заявлений, формам медицинских документов, формам отдела кадров, типовым гражданско-правовым договорам, кредитным договорам.

Настройка задачи сбора образцов цифровых отпечатков в DeviceLock Management Console Новейшая версия DeviceLock DLP 8.3 позволяет снимать цифровые отпечатки с образцов конфиденциальных документов и типовых форм для сбора персональных данных, сохраняя их в собственной базе данных отпечатков, а затем сравнивать их с отпечатками защищаемых при передаче документов. Если процент соответствия отпечатков превышает требуемый порог в соответствии с заданной DLP-политикой, проверяемые документы считаются конфиденциальными и к ним в режиме реального времени применяются заданные в DLP-политике действия — блокировка передачи документа, создание теневой копии, отправка тревожного оповещения в службу ИБ. Рисунок 3. Создание правила проверки содержимого с помощью цифровых отпечатков в DeviceLock Management Console Точность детектирования персональных данных в DeviceLock DLP может быть значительно повышена благодаря возможности создания правил контентной фильтрации с условиями детектирования в сколь угодно сложных комбинациях с использованием логических функций и сочетанием метода цифровых отпечатков с анализом по ключевым словам и регулярным выражениям.

«Самое страшное, что может произойти — изменение юридического механизма вменения в пользу алгоритмов. Это намного страшнее „большого брата“, — считает Гаазе. Но, по мнению Хачуяна, такого никогда не будет. По словам эксперта, говоря о больших данных, важно разделять слежку со стороны спецслужб и реалии новой экономики, когда мы платим Instagram за бесплатный сервис теми действиями, которые совершаем внутри сети, тем самым позволяя ему зарабатывать на продаже наших данных и рекламы. Не менее важно разделять управленческие решения, принимаемые государством, и алгоритмы, помогающие ему. „Если завтра правительство захочет лишить всю молодежь, которая сходила на митинг, возможности получать бесплатное образование, если примет такое управленческое решение, оно исполнит его с большими данными и без. Хотя, конечно, нельзя отрицать, что с алгоритмами дело пойдет куда быстрее“, — заметил он. По словам экспертов, большой объем информации машины получают, анализируя задний план на фотографиях, где могут быть не только здания, но и другие люди. А про детей они сегодня знают едва ли не больше, чем про взрослых. Уже к пятилетнему возрасту в Сети накапливается огромный массив персональных данных о них благодаря тому, что выкладывают родители — начиная со снимков УЗИ и заканчивая постами, фотографиями и видео, заметила в свою очередь культуролог, доцент Шанинки и креативный директор Фонда Егора Гайдара Оксана Мороз. То есть — очень большой объем информации о человеке алгоритмы собирают тогда, когда он сам еще не в состоянии ею поделиться, или это не входило в его планы. По словам Хачуяна, в некоторых странах законодательство уже запрещает распространять данные о детях даже родителям. „В России родитель может дать и передать права на обработку персональных данных несовершеннолетнего, а в Европе нет. То есть, там ты своего ‚годовасика‘ не можешь заинстаграмить, как бы он тебе не нравился, и наверное это хорошо“, — заметил он. Вне поля зрения машин остается лишь старшее поколение, которое практически не оцифровано. Их не просто нет в онлайне. Они жили в доцифровую эпоху, когда все их действия еще не могли быть записаны, поэтому все, что у нас есть — их воспоминания, которые по большому счету представляют собой их субъективное мнение. Однако, по словам Хачуяна, с точки зрения создания профиля конкретного человека лишь 30-40% составляют данные о нем. Остальные 60% — данные его окружения. „Если мне с точки зрения алгоритма не хватит данных о вас, я также могу использовать данные вашего цифрового окружения. Хотя, действительно, старшее поколение имеет в этой истории бесспорное преимущество, особенно люди 60+, про которых практически ничего в онлайне нет, только внуки, которые с бабушкой сфотографировались“, — отметил он. Возможности алгоритмов куда шире, чем можно себе представить, считает Гаазе. „Мы всегда относились к несбывшимся возможностям нашей жизни как к чему-то, чего не существует. Большие данные как алгоритм, структурирующий и описывающий все на свете, позволяет видеть реальность совершенно по-другому. Видеть не только то, что случилось, а все дерево решений целиком“, — отметил эксперт. По его словам, теперь социальная жизнь — это не совокупность сбывшихся маршрутов, а совокупность всех возможных маршрутов вообще, которые определенным образом записаны, зафиксированы и сохранены. „Авария 30 лет назад — это нелепое стечение обстоятельств. Авария сейчас — как возможность — уже из некоторой точки видна. Ее вероятность может быть высчитана и более того — может быть сохранена“, — пояснил социолог. Поэтому социология цифровой эпохи — социология чего-то принципиально нового, считает он. Вместе с тем, большие данные оставляют широкое поле для ошибок, отметил Хачуян. „Раньше социолог, условно, исследовал какие-то определенные параметры. Сейчас он может взять 20 тысяч различных параметров — от цвета попкорна, который предпочитает человек, до цвета волос его подписчиков, и найти между этим математическую зависимость. Но при этом никакой ‚содержательной‘ зависимости между этими фактами не будет“, — отметил он. Здесь важно понимать, что алгоритм — это всегда только то, что вы в него заложили, будь то статистическая выборка или какое-то ваше личное видение, подчеркнул Хачуян. И он в любом случае получится не универсальным решением, а отражением поставленных задач. „Я часто слышу истории в духе: мы решили внедрить большие данные в HR и оказалось, что алгоритм подбирает нам только белых 35-летних мужчин с высшим образованием. По факту же алгоритм построен на основании выборки этой компании, которая таких людей и берет на работу“, — отметил эксперт. По словам Гаазе, Сноуден рассказал нам ровно про это. „ЦРУ написало алгоритм, который вычислял оперативников Аль-Каиды (террористическая организация, запрещена в России) в Пакистане по телефонным звонкам, действиям в соцсетях и геотегам. Семь из десяти оказались журналистами Аль-Джазиры и других СМИ, которые чаще других бывают в зоне боев, у которых оттуда шли звонки, у которых есть переписка с подозрительными почтовыми адресами, откуда они получали заявления террористов“, — отметил он. Поэтому не нужно относиться к большим данным как к „большому брату“, считает Гаазе. „Это, конечно, ‚тварь‘, наделенная некоторой долей личности. Она по крайней мере может сделать свою копию в какой-то промежуток времени, и потом сравнить нынешнюю версию с тем, чем оно было. Но это не ‚большой брат‘. Это скорее очень странный ребенок, который принципиально последователен, выполняя то, что вы от него просите. В этом смысле он не является ни злом, ни благом“, — полагает эксперт. Но вместе с тем, придется признать, что оператор селекции теперь никакая не судьба, и всеми совпадениями управляют алгоритмы, отмечает Гаазе. „Благо, что человек хочет пожениться? Наверное. Благо, что жену ему подберет алгоритм? Не знаю. Я просто говорю, что теперь это возможно. А завтра будет возможно еще лучше, потому что мы выделяем феромоны, которые содержат ДНК, и алгоритм сможет подобрать идеального генетического партнера. Я не луддит, я просто пытаюсь обозначить мгновенную экспансию границ возможного в этом мире“, — отметил Гаазе. По мнению эксперта, новое общество предполагает возникновение новой этики. „Должно быть выдвинуто очень простое политическое требование, состоящее из двух пунктов, связанных друг с другом. Первое — не должно быть в работе с вашими личными данными закрытых от вас алгоритмов. Второе — не должно быть алгоритмов, принцип работы которых вы не понимаете. Все, что проходит это требование, — благо“, — считает он. По словам Гаазе, один из крупнейших в мире инвестиционных банков Goldman Sachs, который в свое время скупил 30% неликвидного жилья в США и сдает его в аренду, написал алгоритм, который исходя из ваших постов в соцсетях и темпов выплаты по аренде, может расторгнуть с вами договор (это записано в нем). И я хочу знать, как работает алгоритм, который оценивает мое поведение в соцсетях, если в результате я могу лишиться жилья», — отметил он. «Кстати, в Америке есть закон, обязывающий банк, если он использует алгоритмы при выдаче кредитов, объяснить — мы тебе кредит не дадим, потому что ты 20 раз в сториз опубликовала, как пьешь алкоголь. А вот в России такого закона нет», — заметил Хачуян. По его словам, в разных странах принято законодательство о защите персональных данных, но нет ничего про манипуляцию алгоритмами. «Думаю, в ближайшие пять лет ничего такого и не появится по одной простой причине: такой хороший закон в короткий срок не напишешь, а если напишешь, он защитит людей, но повредит рынку. Алгоритмы Facebook — коммерческая тайна. Несмотря на то, что он такого же размера, как некоторые государства, он не обязан открывать внутренности своих алгоритмов никому, потому что защищен патентом. Соответственно, никто никогда не проверит, как это работает, и почему вы получили именно эту информацию. Государство это понимает. Поэтому вся эта история с регулированием — история не ближайшего будущего», — считает эксперт.