Проблемы выявления утечек конфиденциальной информации в неструктурированных данных

Г.В. Гарбузов

***Аннотация*—В настоящей статье рассматривается проблема выявления утечек конфиденциальной информации, представленной в неструктурированном виде, с использованием современных технологий защиты от утечек. В частности, рассмотрены основные свойства неструктурированных данных и отмечена актуальность совершенствования способов их обработки на современном предприятии. Проведен обзор используемых сегодня технологий защиты от утечек и оценена их эффективность для обработки структурированных и неструктурированных данных. Также предложен подход повышения эффективности технологий защиты от утечек информации в неструктурированных данных. Постановка задачи: выявить основные недостатки существующих технологий защиты от утечки информации в неструктурированных данных и предложить меры по их совершенствованию. Результаты: подтверждена актуальность проблемы обработки неструктурированных данных и выявления в них критичной информации, проведен обзор существующих технологий защиты от утечек, обозначены их недостатки и предложен способ повышения их эффективности при работе с отдельными типами неструктурированных данных. Практическая значимость: предложенные подходы могут использоваться специалистами коммерческих и некоммерческих организаций при проектировании систем информационной безопасности, предназначенных для защиты нематериальных активов, представленных в неструктурированном виде. Обсуждение: представлен способ расширения возможностей классических систем защиты от утечек, основанных на статичных правилах, с использованием технологий искусственного интеллекта.**

***Ключевые слова*— коммерческая тайна, неструктурированные данные, ноу-хау, утечка информации, технологии защиты от утечек информации, Data Leak Protection, защита информации.**

I. Классификация информации как ключевой элемент защиты от утечек

Как было неоднократно отмечено [1], [2], нематериальные активы вообще и информация в частности являются ценным активом современного коммерческого предприятия, который необходимо защищать от обесценивания. Одной из существенных угроз, прямо влияющих на ценность информации, является угроза её конфиденциальности: информация может иметь реальную ценность в силу неизвестности третьим лицам. Реализацию этой угрозы мы называем утечкой информации (т.е. несанкционированное и неконтролируемое обладателем распространение информации) и, говоря о защите нематериальных активов буем иметь в виду именно защиту ценной информации от утечек, а саму информацию – конфиденциальной. Утечка и разглашение такой конфиденциальной информации приводят к обесцениванию нематериального актива.

Для защиты от утечек на различных этапах жизненного цикла [3] могут использоваться различные технологии, но согласно некоторым исследованиям, выполненным на основе анализа публикаций об используемых методах защиты конфиденциальной информации в период с 2011 по 2022 годы [4], для защиты от утечек информации и противодействию внутренним угрозам самым популярным (используемым наибольшим количеством организаций) методом защиты конфиденциальной информации является шифрование (около 40%), а на втором месте использование технологий машинного обучения (около 12%), применяемых в дополнение в традиционным DLP (Data Loss Prevention) технологиям. Но независимо от того, какую технологию выбирает организация, она должна применяться точечно к выбранным объектам защиты и базироваться на а) результатах классификации информации и б) способности качественно её идентифицировать (обнаруживать) в файловых ресурсах или каналах связи. В случае игнорирования или некачественного выполнения классификации, критичная информация будет неверно идентифицирована, что в конечном итоге приведет либо к её утечке (недостаток контроля), либо к необоснованным усложнениям бизнес процессов (избыток контроля), что в одинаковой степени может негативно отразиться на прибыли коммерческой организации.

Результаты классификации на основе разработанных политик и правил классификации информации должны использоваться на всех этапах жизненного цикла информации: от создания, предполагающего принятие решения о критичности информации и внесения специальной маркировки, до контроля соблюдения политик безопасности, связанного с непосредственной

Статья получена 16 января 2025.

Георгий Валерьевич Гарбузов – аспирант кафедры информационной безопасности факультета информационных технологий и анализа больших данных, Финансовый университет при Правительстве Российской Федерации; ORCID: http://orcid.org/0009-0008-7717-1488 (e-mail: g.garbuzov@mail.ru)

идентификацией (выявлением) критичной информации в общем объеме хранимых или передаваемых данных. Руководствуясь данным выше определением утечки информации, можно сделать вывод о том, что противодействие утечкам прямо связано с технологиями детектирования классифицированной информации в потоках данных, передаваемых из организации вовне, и оценке соответствия конкретной передачи (состав данных, оснований для передачи, способа передачи, маршрута передачи) политикам безопасности. В случае выявления нарушений, передача информации может быть заблокирована, причем блокировка может быть верной или ошибочной (ложноположительные срабатывания, т.н. ошибка первого рода), здесь же имеет смысл говорить о ложноотрицательных срабатываниях, когда передача нарушала требования безопасности, но заблокирована не была (система восприняла отправку как легитимную, ошибка второго рода). Для выявления подобных нарушений сегодня используются упомянутые выше системы класса DLP, которые неплохо работают со структурированными данными, но имеют ряд серьезных проблем при обработке информации в неструктурированном виде, мы обсудим их ниже.

II. Проблемы выявления утечек неструктурированной информации

Информация, представляющая ценность для организации, может быть представлена в структурированном и неструктурированном виде, рассмотрим их подробнее.

Структурированная информация представлена в виде строк и колонок, организованных в базы данных согласно единым правилам, называемым моделью данных. Элементами в таком структурированном наборе могут быть введенные людьми имена, адреса, даты и номера кредитных карт, а также объективные данные: события сетевых журналов, данные электронной коммерции и др. Организация информации в структурированном виде имеет ряд полезных свойств: такую информацию просто создавать и считывать, хранение информации в структурированном виде требует меньше ресурсов. Кроме того, с такой информацией проще работать (осуществлять поиск и извлекать нужные данные) и её проще защищать традиционными способами (например, используя встроенные механизмы управления доступом и контроля запросов систем управления базами данных).

Неструктурированная информация, согласно Gartner[1], сегодня составляют от 80 до 90% всей информации, которой располагает современная организация, а её объемы растут втрое быстрее, чем объемы структурированной информации. Эта информация не организована в соответствии с заранее определенной моделью или структурой данных, её элементы не имеют четких взаимосвязей, а поэтому она не может храниться в виде базы данных. Неструктурированная информация представлена в виде сообщений электронной почты или отдельных текстовых, графических, аудио- и видео файлах различных форматов. Их невозможно обработать методами, применяемыми для анализа структурированной информации, поскольку ценность такой информации определяется не позицией ячейки в базе данных или атрибутом модели данных, а её содержанием, предполагающим её осмысление. Также серьезно затрудняет обработку неструктурированной информации её значительные объёмы и высокая изменчивость, поскольку эти данные создаются в реальном времени различными источниками, отражая деловой ритм организации. Вместе с тем, именно неструктурированная информация, как правило, является нематериальным активом организации. Примерами могут служить производственные и лабораторные регламенты, описания изобретений и способов модернизации производства, патентные заявки до регистрации права на изобретение, файлы с презентациями стратегических планов, исследовательские отчеты и др.

Ярким примером ценной неструктурированной информации является информация, составляющая коммерческую тайну и секреты производства. Федеральный закон [5] описывает информацию, составляющую коммерческую тайну, как «сведения любого характера (производственные, технические, экономические, организационные и другие), в том числе о результатах интеллектуальной деятельности в научно-технической сфере, а также сведения о способах осуществления профессиональной деятельности, которые имеют действительную или потенциальную коммерческую ценность в силу неизвестности их третьим лицам…», а [6] вводит понятие секрета производства (ноу-хау), которым признаются «сведения любого характера (производственные, технические, экономические, организационные и другие) о результатах интеллектуальной деятельности в научно-технической сфере и о способах осуществления профессиональной деятельности, имеющие действительную или потенциальную коммерческую ценность вследствие неизвестности их третьим лицам, если к таким сведениям у третьих лиц нет свободного доступа на законном основании и обладатель таких сведений принимает разумные меры для соблюдения их конфиденциальности, в том числе путем введения режима коммерческой тайны».

Важно отметить, что в обоих случаях необходимым условием является принятие обладателем мер для соблюдения конфиденциальности указанной информации. Если вспомнить указанные выше оценки (от 80 до 90% от общего объема информации в организации – неструктурированные данные) и тенденции (рост утечек информации, составляющей коммерческую тайну и интеллектуальную собственность в три раза в 2023 году по отношению к 2022 году[2]), можно сделать вывод о том, что упомянутая ранее задача классификации и идентификации конфиденциальной информации, хранящейся и передаваемой именно в

[1] Consult the Board: Unstructured Data Management // Gartner. 18 мая 2023. URL: https://www.gartner.com/en/documents/4373899

[2] Отчет об оценке ущерба вследствие утечек информации // Экспертно-аналитический центр ГК InfoWatch. 06 сентября 2023. URL: https://www.infowatch.ru/analytics/analitika/otsenka-uscherba-vsledstvie-utechek-informatsii

неструктурированном виде, выходит на первый план.

III. Технологии выявления утечек информации в неструктурированных данных

Следует констатировать, что на сегодняшний день не существует средств, позволяющих полноценно обрабатывать неструктурированную информацию, и, следовательно, эффективно защищаться от её утечек, а системы класса DCAP (Data-Centric Audit and Protection), DAG или уже упомянутые DLP, которые могли бы быть использованы при решении этих задач, имеют в своем арсенале лишь ограниченный набор технологий. DAG и DCAP системы эффективно работают только с уже классифицированными и маркированными данными, а DLP системы в состоянии идентифицировать конфиденциальную информацию в немаркированных данных, но использующиеся технологии несовершенны. Рассмотрим подробнее DLP системы.

DLP системы предназначены [7] для выявления и реагирования на нарушения при хранении и передаче конфиденциальной информации, при этом в DLP системе различают сущности (объекты защиты, на обнаружение которых реагирует DLP система) и политики (действия, которые предпринимает DLP система при обнаружении объекта, например, блокировка передачи информации). Современные DLP системы способны осуществлять анализ различных типов хранилищ и трафика, называемого каналами (электронная почта, файловые ресурсы, http- и ftp-трафик, облачные ресурсы, интернет мессенджеры IP телефония и др.), и контролировать передачу информации на различных носителях (печать, съемные и медиа носители информации), а также буфер обмена рабочей станции. Типичная структурная схема DLP системы приведена на Рис. 1.

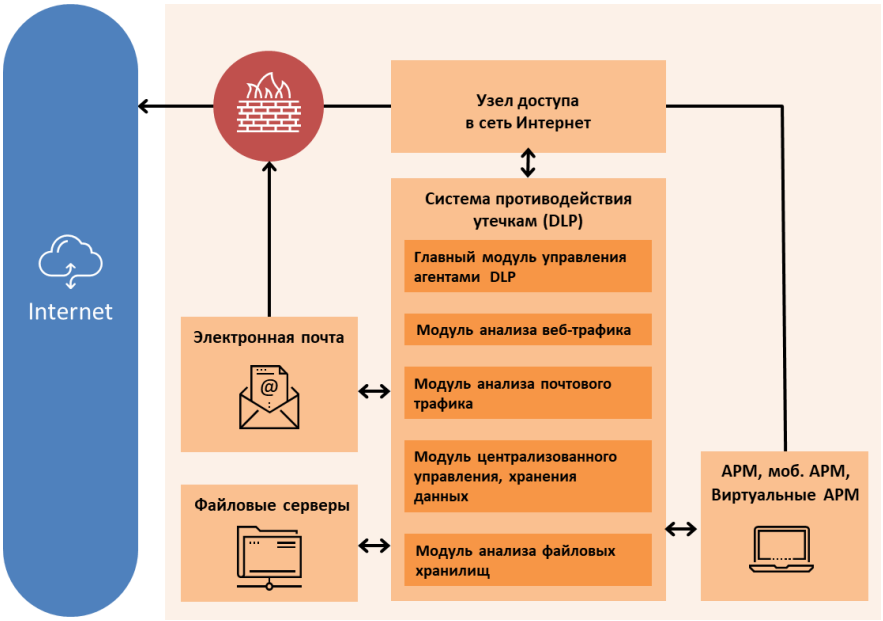

Рис. 1: Структурная схема DLP системы

Функционально на ней можно выделить следующие элементы:

- *Автоматизированные рабочие места (АРМ)*: рабочие станции пользователей, на которых создается и хранится информация, а также передается информация по телекоммуникационным каналам или через внешние порты оборудования;
- *Электронная почта*: система электронной почты на безе MS Exchange или подобного;
- *Узел доступа в Интернет*: как правило, реализован на базе решения класса «Proxy-сервер», позволяющего управлять проходящим через сервер трафиком «на лету»;
- *Файловые серверы*: системы хранения информации, построенные в корпоративной сети (облачные среды к ним не относятся).

Идентификация конфиденциальной информации осуществляется непосредственно в трафике, его копии, в файловом ресурсе или на рабочей станции при помощи отдельных модулей (таких как краулер, предназначенных для сканирования и поиска конфиденциальной информации в файловых ресурсах) или агентов, с использованием следующих основных методов:

1. *Цифровые отпечатки*. В основе метода цифровых отпечатков лежит сопоставление конкретных документов с их цифровыми отпечатками, которые представляют собой результат математического преобразования хэш-функции. Этот результат (хэш) для каждого документа будет разным, поэтому точность идентификации такого документа в хранилище или трафике очень высока, а число ошибок первого и второго рода околонулевым. Метод цифровых отпечатков подходит для защиты как структурированных, так и неструктурированных данных, но применять его повсеместно, для защиты всей информации в организации, чрезвычайно трудозатратно и практически неосуществимо на практике (т.к. один и тот же документ различных редакций будет иметь различные цифровые отпечатки, что вызовет колоссальный рост числа

объектов защиты). На практике метод применяется для точечного контроля отдельных конкретных документов, содержащих критичные данные;

2. *Регулярные выражения*. Метод основан на результатах морфологического анализа и призван выявить закономерности в анализируемом тексте, указывающие на его отношение к какому-либо объекту защиты. Метод подходит для идентификации структурированных данных (выгрузок из баз данных, отдельных записей, сформированных по каким-либо алгоритмам – номера телефонов, ИНН, СНИЛС и т.д.). Для анализа неструктурированной информации не подходит;
3. *Лингвистический анализ*. Метод основан на анализе текста с использованием предустановленных словарей или отдельных слов. Подходит для идентификации маркированной (снабженной грифом конфиденциальности) и немаркированной конфиденциальной информации, в том числе неструктурированной. В случае идентификации немаркированной информации, если ключевые слова присутствуют в общих словарях (например, «совет директоров», «стратегия развития» и т.д.), сопровождается значительным количеством ошибок первого рода (до 95%), на разбор которых расходуется значительное время и экспертизы офицеров безопасности. Улучшить ситуацию возможно с использованием узкоспециальных ключевых слов и выражений, присущих именно конфиденциальным документам (например, «фармакопейная статья предприятия» или «расчет по антигенам» из отрасли фармацевтики), но для подавляющего числа организаций сферы электронной коммерции, ИТ, медиа и других такая возможность будет либо вовсе недоступна, либо не сможет покрыть всех случаев, в которых требуется идентификация конфиденциальной информации. В целом, использование метода целесообразно только при работе с неструктурированной конфиденциальной информацией, снабженной маркировкой (грифом или пометкой) конфиденциальности, что требует высокой культуры работы с конфиденциальными документами и сильно сужает область применения в коммерческих организациях;
4. *Специальные технологии*. Методы, основанные на специальных и проприетарных алгоритмах, например, алгоритме Луна[3] для идентификации в потоке данных номеров пластиковых карт. Не исключая ошибки первого и второго рода, как правило, применяется в узких и специальных областях для идентификации конкретных структурированных данных. Для анализа неструктурированной информации не подходит.

Как можно заметить, ни один из используемых на сегодня методов, используемых в DLP системах, не обеспечивает требуемого уровня точности и полноты идентификации конфиденциальной информации в неструктурированных данных, следовательно, с точки зрения кардинального повышения качества процесса защиты от утечек неструктурированной конфиденциальной информации перед организациями остро встает необходимость привлечения новых технологий. Самым перспективным путем видится использование технологий искусственного интеллекта, которые смогут работать в составе, например, DLP систем и расширить возможности используемых методов.

История возникновения искусственного интеллекта как концепции, согласно некоторым источникам [8, 9], ведет свое начало с 1955 года. Независимо от способа реализации, которые претерпели значительные изменения в сравнении с первыми прототипами, принцип функционирования систем искусственного интеллекта неизменен и означает способность предсказывать будущие состояния (события) на основе уроков, извлеченных из анализируемых данных, что подразумевает гибкую адаптацию (т.н. обучаемость системы).

В России и мире в настоящее время этот вопрос активно прорабатывается. Например, в [10] автор анализирует возможности применения нейронных сетей в DLP системах с нескольких аспектов, а также отмечает возможности и перспективы дальнейшего развития применения технологий искусственного интеллекта в DLP системах. Однако, данный материал носит чисто теоретический характер и не подкреплен никакими эмпирическими данными. В другой работе [11] автор отмечает не только возможности, но и риски применения искусственного интеллекта и рассматривает аспекты использования т.н. доверенного искусственного интеллекта (Trusted Artificial Intelligence), который сам может стать объектом атаки на конфиденциальность и нуждается в защите. При этом в заключении автор признает, что «представленные технологические стратегии — не более чем теоретизация на тему планирующихся в ближайшем будущем процессов».

В зарубежных исследованиях напротив, приводится конкретный опыт применения новых технологий при решении задач защиты от утечек. Например, в [12, С. 143-144] авторы исследовали возможность применения технологий искусственного интеллекта при решении конкретных задач и описали разработанный ими программный прототип, который собирает данные о сети и информационных потоках, выводя обобщенные данные об уровнях риска для конкретных информационных активах на дашборд. В статье описаны ключевые элементы (риск, актив, угроза) прототипа и методология его создания, но процесс создания (обучения) модели идентификации конфиденциальной информации описан чрезвычайно сжато, в то время как качество (релевантность) модели является решающим фактором и, в свою очередь, прямо зависит от качества набора данных (датасета), используемого для обучения. Подготовка датасета для обучения модели искусственного интеллекта (особенно когда речь идет об

[3] Алгоритм Луна (англ. Luhn algorithm) — алгоритм вычисления контрольной цифры номера пластиковой карты в соответствии со стандартом ISO/IEC 7812, разработан инженером IBM Хансом Питером Луном (Hans Peter Luhn) в 1954 году.

обучении модели для идентификации критичной информации) представляется отдельной непростой задачей и важной областью исследования, которой посвящен ряд интересных публикаций. Например, в работе [13] авторами предлагается методология обучения модели на основе датасета, сформированного путем выделения из текста именованных сущностей (Named Entities), ассоциированных с конфиденциальной информацией, и применения методов математической статистики. Модель была апробирована на 1000 документов из отраслей права и медицины, и показала удовлетворительные результаты: точность 0,79 и 0,75 (соответственно), полнота 0,63 и 0,53 (соответственно). Данная методология не может быть рекомендована для идентификации критичных данных, но может использоваться при обучении массовых и общеотраслевых моделей, т.к. помогает значительно сократить объем привлечения к подготовке датасетов квалифицированных экспертов.

В ряде других публикаций [14-16] исследуются вопросы применения технологий искусственного интеллекта для противодействия различным угрозам безопасности и все они сходятся в одном – использование машинного обучения способно в значительной степени помочь как в идентификации угрозы, так и в выборе стратегии реагирования на неё. С точки зрения защиты от утечек конфиденциальной информации использование технологий искусственного интеллекта даст возможность не только идентифицировать конфиденциальную информацию в конкретном неструктурированном тексте (например, в составе или отдельном модуле DLP системы), но и использовать автоэнкодеры, предназначенные для упорядочения неструктурированной информации и преобразования её в структурированную с последующим применением традиционных методов выявления утечек информации, использующихся в DLP и описанных ранее.

IV. Заключение

Задача классификации и идентификации критичной информации является чрезвычайно актуальной для коммерческих организаций, поскольку является неотъемлемой частью эффективного процесса защиты от её утечек. Однако, если учесть, что значительная часть ценной информации представлена в неструктурированном виде, организации необходимо для защиты от угрозы утечки применять технологии, доказавшие свою эффективность при работе именно с неструктурированными данными. Технологии, реализованные в современных системах защиты от утечек, хорошо справляются с анализом структурированных данных, но для работы с неструктурированной информацией неэффективны. Учитывая объемы неструктурированной информации, скорость её накопления и изменчивость, эти технологии должны позволять гибко и оперативно управлять политиками безопасности, направленными на предотвращение возможных утечек, с минимальным участием человека. В качестве такой технологии предлагается исследовать и апробировать возможность применения технологий машинного обучения и искусственного интеллекта, разработав модель, которая может быть использована в составе существующих систем безопасности (DLP, DAG, DCAP и др.) и основанный на использовании этой модели метод защиты от утечек конфиденциальной информации.

Данные проблемы определяют необходимость разработки как методологии решения указанных задач, так и техническую целесообразность разработки средств автоматизации, апробированных при работе, прежде всего, с неструктурированной информацией. Такая постановка научно-технической проблемы, связана с низкой эффективностью существующих средств защиты от утечек информации при работе с коммерческой тайной и интеллектуальной собственностью (что, вероятно, влияет на кратный рост объемов их утечек в последнее время), с одной стороны, и со значительным объемом и стоимостью указанной информации в структуре экономики современной коммерческой организации, с другой.

Возможности идентификации произвольной конфиденциальной информации в неструктурированных данных в полной мере обеспечиваются методами машинного обучения и искусственного интеллекта. Однако, опираясь на указанные технологии, мы привносим в традиционный процесс защиты от утечек конфиденциальной информации новые проблемы, присущие задачам машинного обучения, поэтому актуальной является задача разработки и апробации методики подготовки набора данных (датасета), позволяющей быстро адаптировать обученную на нем модель к изменяющимся условиям, в которых существует современная коммерческая организация. Отдельную сложность здесь представляет сбор реальной критичной информации для формирования релевантного набора данных. По этой практически важной задаче в литературе найдены только формулировки целей, но ни одного содержательного результата.

Библиография

[1] Гарбузов Г. В. Проблемы дефиниций и постановки целей защиты от утечек информации ограниченного доступа // International Journal of Open Information Technologies. 2024. Т. 12, № 5. С. 185-191. EDN: ZXVIYQ

[2] Ferrara E. Determine The Business Value Of An Effective Security Program – Information Security Economics 101. Forrester Research, Inc., October 2, 2012.

[3] Гарбузов Г. В. Технологии защиты нематериальных активов от атак на конфиденциальность // International Journal of Open Information Technologies. 2024. Т. 12, № 9. С. 142-149. EDN: CXXTYY

[4] Herrera Montano I., García Aranda J.J., Ramos Diaz J., et al. Survey of Techniques on Data Leakage Protection and Methods to address the Insider threat // Cluster Computing. 2022. Vol. 25. P. 4289-4302. doi: https://doi.org/10.1007/s10586-022-03668-2

[5] О коммерческой тайне : федер. закон от 29.07.2004 г. № 98-ФЗ (последняя редакция) : принят Государственной Думой 9 июля 2004 г. URL: https://www.consultant.ru/document/cons_doc_LAW_48699/

[6] Гражданский Кодекс РФ. Часть четвертая : федер. закон от 18 декабря 2006 г. № 230-ФЗ : принят Государственной Думой 24 ноября 2006 г. URL: https://www.consultant.ru/document/cons_doc_LAW_64629

[7] Зарубин А. В., Смирнов М. Б., Харитонов С. В., Денисов Д. В. Основные драйверы и тенденции развития DLP-систем в

Российской Федерации // Прикладная информатика. 2020. Т. 15. № 3. С. 75-90. doi: https://doi.org/10.37791/2687-0649-2020-15-3-75-90

[8] Haenlein M., Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence // California Management Review. 2019. Vol. 61, No. 4. P. 5-14. doi: https://doi.org/10.1177/0008125619864925 7

[9] Wei J. Research progress and application of computer artificial intelligence technology // MATEC Web of Conferences. 2018. Vol. 176. Article number: 01043. doi: https://doi.org/10.1051/matecconf/201817601043

[10] Артюшкина Е. С., Скакун О. О., Гузь А. Р. Использование искусственного интеллекта в DLP-системах // Прикладные экономические исследования. 2023. № 2. С. 123-129. doi: https://doi.org/10.47576/2949- 1908_2023_2_123

[11] Авдошин С. М., Песоцкая Е. Ю. Доверенный искусственный интеллект как способ цифровой защиты // Бизнес-информатика. 2022. Т. 16, № 2. С. 62-73. doi: https://doi.org/10.17323/2587-814X.2022.2. 62.73

[12] Donglan Liu, Xin Liu, Lei Ma, Yingxian Chang, Rui Wang, Hao Zhang, Hao Yu, Wenting Wang. Research on Leakage Prevention Technology of Sensitive Data based on Artificial Intelligence // 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC). Beijing, China: IEEE Computer Society, 2020. P. 142-145. doi: https://doi.org/10.1109/ICEIEC49280.2020.9152286

[13] Martinelli F., Marulli F., Mercaldo F., Marrone S., Santone A. Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence // 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK: IEEE Computer Society, 2020. P. 1-8. doi: https://doi.org/10.1109/IJCNN48605.2020.9206801

[14] Kim J., Lee C., Chang H. The Development of a Security Evaluation Model Focused on Information Leakage Protection for Sustainable Growth // Sustainability. 2020. Vol. 12, issue 24. Article number: 10639. https://doi.org/10.3390/su122410639

[15] Zhu T., Ye D., Wang W., Zhou W., Yu P.S. More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence // IEEE Transactions on Knowledge and Data Engineering. 2022. Vol. 34, No. 6. P. 2824-2843. doi: https://doi.org/10.1109/TKDE.2020.3014246

[16] Guha A., Samanta D., Banerjee A., Agarwal D. Deep Learning Model for Information Loss Prevention From Multi-Page Digital Documents // IEEE Access. 2021. Vol. 9. P. 80451-80465. doi: https://doi.org/10.1109/ACCESS.2021.3084841

Issues in Detecting Confidential Information Leaks in Unstructured Data

Georgy Garbuzov

***Abstract*— this article addresses the problem detection of leaks of confidential information, presented in the unstructured form, using modern technologies of protection against leaks. In particular, the main properties of unstructured data are considered and the relevance of improving the methods of their processing at the modern commercial enterprise. The review of leakage protection technologies used today is carried out and their efficiency for processing of structured data is evaluated.**

An approach to improve the effectiveness of information leakage protection technologies in unstructured data is also proposed. Problem statement: to identify the main drawbacks of existing technologies of protection against information leakage in unstructured data and propose measures for their improvement. Main results: The relevance of the problem of unstructured data processing and identification of critical information in them was confirmed, the existing leakage protection technologies were reviewed, their disadvantages were outlined and a way to improve their efficiency when working with certain types of unstructured data was proposed. Practical significance: The proposed approaches can be used by specialists of commercial and non-commercial organizations when designing information security systems designed to protect intangible assets represented in unstructured form. Discussion: A method of extending the capabilities of classical leakage protection systems based on static rules using artificial intelligence technologies is presented

***Keywords* - trade secret, unstructured data, know-how, information leakage, information leakage protection technologies, Data Leak Protection, information protection.**

REFERENCES

[1] Garbuzov G. Problems of Definitions and Setting Goals for Data Leaks Protection. *International Journal of Open Information Technologies*, 2024, vol. 12, no. 5, pp. 185-191. (In Russ., abstract in Eng.) EDN: ZXVIYQ

[2] Ferrara E. Determine The Business Value Of An Effective Security Program – Information Security Economics 101. Forrester Research, Inc., October 2, 2012.

[3] Garbuzov G. Technologies for Protecting Intangible Assets from Confidentiality Attacks. *International Journal of Open Information Technologies*, 2024, vol. 12, no. 9, pp. 142-149. (In Russ., abstract in Eng.) EDN: CXXTYY

[4] Herrera Montano I., García Aranda J.J., Ramos Diaz J., et al. Survey of Techniques on Data Leakage Protection and Methods to address the Insider threat. *Cluster Computing*, 2022, vol. 25, pp. 4289-4302. doi: https://doi.org/10.1007/s10586-022-03668-2

[5] [On Commercial Secrecy (with The Amendments and Additions): Federal Law No. 98-FZ of July 29, 2004: Adopted by the State Duma on 9 July, 2004]. [Online]. Available: https://www.consultant.ru/document/cons_doc_LAW_48699 (In Russ.)

[6] Civil Code of the Russian Federation (Part Four, (with The Amendments and Additions): Federal Law No. 230-FZ of December 18, 2006: Adopted by the State Duma on November 24, 2006]. [Online]. Available: https://www.consultant.ru/document/cons_doc_LAW_64629/ (In Russ.)

[7] Zarubin A., Smirnov B., Kharitonov S., Denisov D., Main drivers and trends of DLP systems development in the Russian Federation. *Prikladnaya informatika* = Journal of Applied Informatics, 2020, vol. 15, no. 3, pp. 75-90. (In Russ., abstract in Eng.) doi: https://doi.org/10.37791/2687-0649-2020-15-3-75-90

[8] Haenlein M., Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 2019, vol. 61, no. 4, pp. 5-14. doi: https://doi.org/10.1177/0008125619864925 7

[9] Wei J. Research progress and application of computer artificial intelligence technology. *MATEC Web of Conferences*, 2018, vol. 176. Article number: 01043. doi: https://doi.org/10.1051/matecconf/201817601043

[10] Artyushkina E.S., Skakun O.O., Guz A.R. Using artificial intelligence in DLP systems. *Applied economic research*, 2023, no. 2, pp. 123-129. (In Russ., abstract in Eng.) doi: https://doi.org/10.47576/2949-1908_2023_2_123

[11] Avdoshin S.M., Pesotskaya E.Yu. Trusted artificial intelligence: Strengthening digital protection. *Business Informatics*, 2022, vol. 16, no. 2, pp. 62-73. doi: https://doi.org/10.17323/2587-814X.2022.2.62.73

[12] Donglan Liu, Xin Liu, Lei Ma, Yingxian Chang, Rui Wang, Hao Zhang, Hao Yu, Wenting Wang. Research on Leakage Prevention Technology of Sensitive Data based on Artificial Intelligence. In: *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC).* Beijing, China: IEEE Computer Society; 2020. pp. 142-145. doi: https://doi.org/10.1109/ICEIEC49280.2020.9152286

[13] Martinelli F., Marulli F., Mercaldo F., Marrone S., Santone A. Enhanced Privacy and Data Protection using Natural Language Processing and Artificial Intelligence. In: *2020 International Joint Conference on Neural Networks (IJCNN).* Glasgow, UK: IEEE Computer Society, 2020. pp. 1-8. doi: https://doi.org/10.1109/IJCNN48605.2020.9206801

[14] Kim J., Lee C., Chang H. The Development of a Security Evaluation Model Focused on Information Leakage Protection for Sustainable Growth. *Sustainability*, 2020, vol. 12, issue 24. Article number: 10639. https://doi.org/10.3390/su122410639

[15] Zhu T., Ye D., Wang W., Zhou W., Yu P.S. More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence. *IEEE Transactions on Knowledge and Data Engineering*, 2022, vol. 34, no. 6, pp. 2824-2843. doi: https://doi.org/10.1109/TKDE.2020.3014246

[16] Guha A., Samanta D., Banerjee A., Agarwal D. Deep Learning Model for Information Loss Prevention From Multi-Page Digital Documents. *IEEE Access*, 2021, vol. 9, pp. 80451-80465. doi: https://doi.org/10.1109/ACCESS.2021.3084841

About the authors:

Georgy Garbuzov, Postgraduate Student, Financial University under the Government of the Russian Federation, ORCID: http://orcid.org/0009-0008-7717-1488 (e-mail: g.garbuzov@mail.ru)