III. Проблемы обеспечения информационной безопасности

Мещеряков Р.В., Евсютин О.О., Исхакова А.О., Душкин А.В.

Обеспечение информационной безопасности слабоструктурированной информации при решении задач защиты информации

Аннотация: В работе обсуждается вопрос реализации защиты информации в системах обработки слабоструктурированных данных. Приводятся особенности инфраструктуры и обработки, влияющие на моделирование подсистемы информационной безопасности. Перечисляются перспективы развития направления и возможные подходы к решению задач.

Ключевые слова: слабоструктурированная информация, обработка данных, мониторинг информационной безопасности, защита данных, безопасность знаний, риски информационной безопасности

1. Обработка слабоструктурированных данных

Слабоструктурированными принято называть данные, не имеющие фиксированного формата и четкой структуры. Такие информационные массивы могут быть более гибкими по сравнению со строго структурированными данными, например, данными в реляционных базах данных, что делает их привлекательными в решении сложных задач, выявлению новых знаний. Однако в отличие от неструктурированных данных, слабоструктурированные данные все же содержат некоторую организацию или метаданные, которые помогают извлекать из них информацию.

Современное развитие технологий позволяет обрабатывать слабоструктурированную информацию в объемах, которые ранее были недоступны. Это провоцирует развитие соответствующих систем обработки и, соответственно, требует создания новых методов оценки информационной безопасности таких систем [1-2].

Особенностью обработки такой информации является ее неоднородность. Слабоструктурированные данные могут быть представлены в различных форматах, что затрудняет их обработку и анализ. Это требует разработки универсальных подходов для работы с разными типами данных. Содержимое слабоструктурированных данных может быть нечетким и многозначным, что затрудняет автоматическое извлечение информации [3]. Например, одно и то же слово может иметь разные значения в зависимости от контекста.

Машинное обучение и глубокое обучение позволяют эффективно обрабатывать данные, не имеющие четкой структуры, обучая алгоритмы на больших объемах информации. Эти алгоритмы позволяют выделить ключевые признаки из данных, которые могут помочь в дальнейшем анализе и принятии решений [4-5]. Например, в текстах это могут быть ключевые слова или фразы, а в изображениях – контуры и текстуры.

Перечисленные особенности определяют подходы и инструменты, которые используются в области обработки слабоструктурированных данных, и требуют постоянного обновления знаний и адаптации методов к новым вызовам.

2. Искажения и их коррекция в слабоструктурированных данных

Наличие искажений в текстах значительно снижает эффективность их автоматической обработки. Анализ и коррекция искаженных текстов актуальны в следующих областях: коррекция текстов, набранных с ошибками, обработка коротких сообщений в социальных сетях, оптическое распознавание символов, распознавание речи, распознавание рукописного текста, машинный перевод, добывание информации из текста, обработка текстовых запросов, классификация и аннотирование текстов, а также задачи анализа оперативной обстановки с помощью систем дополненной реальности, использующих перевод речи и изображений документов с иностранных языков.

Под сильным искажением понимается уровень, который близок или превосходит теоретико-информационную границу. На сегодняшний день известно, что при сильных искажениях невозможно в среднем однозначное восстановление текста. Однако

отдельные фрагменты текста и оценки, характеризующих текст признаков, могут быть получены.

В телекоммуникационных каналах связи, как правило, используются помехоустойчивое кодирование, а проводя параллель – информация на естественном языке (в текстовом, визуальном и акустическом сигнальном виде) имеет избыточность и позволяет извлекать существенную информацию и исправлять ошибки. Известно, что словесная разборчивость речи имеет большее значение по сравнению со слоговой при восприятии речи человеком. Человек в этом случае выступает системой обнаружения и коррекции ошибок, однако увеличивающийся поток информации в различных каналах не позволяет всегда использовать человека и возникает необходимость создания методологического аппарата для создания автоматических систем.

Распространенные программные средства коррекции, хорошо работающие при малых искажениях в тексте, в случае текстов с высоким уровнем искажений, вне зависимости от их происхождения (набранных с ошибками на клавиатуре, полученных в результате распознавания речи в условиях шумов и др.), показывают неудовлетворительные результаты. Это делает необходимым разработку качественно новых самостоятельных подходов к коррекции сильно искаженных текстов.

Решение проблемы восстановления сильно искаженных текстов является актуальным также для задач оценивания информативности возможных каналов утечки информации.

3. Защита слабоструктурированных данных

Начальным этапом построения практически любой системы защиты информации является проведение аудита информационных ресурсов. Известны подходы по выявлению и классификации информационных активов, позволяющих ранжировать их важность, а в некоторых случаях и критичность. Указанные подходы являются известными и хорошо проработанными, например, в государственных информационных системах, системах обработки персональных данных, на объектах критической информационной инфраструктуры и других системах, обрабатывающих информацию, доступ к которой законодательно и с использованием нормативно-правовых актов ограничен. Вместе с тем наличие в системе

слабоструктурированной информации (в том числе многомодальной информации – речь, изображения, текст) приводит к понятию безопасности знаний, т.к. на основе такой информации можно получить новую информацию. Таким образом, методы по оценке стоимости, критичности и других свойств претерпевают изменения и требуют совершенствования.

Подходы к построению защиты информации в системах, которые обрабатывают структурированную информацию, широко известны. В зависимости от этапа жизненного цикла и доступности ресурсов могут быть использованы как методы архитектурного встраивания в систему обработки информации, так и наложенные средства. Современный подход Secure-By-Design решает ряд архитектурных проблем, однако не всегда может быть использован при проектировании системы защиты слабоструктурированной информации.

При функционировании систем обработки слабоструктурированной информации требуется регулярный мониторинг – в дополнение к обеспечению конфиденциальности, целостности и доступности сохранение иных свойств информации.

Наглядным примером такой сложной системы является использованием методов стеганографии в обеспечении защиты информации при передаче по каналам связи.

Методы цифровой стеганографии направлены на обеспечение конфиденциальности информации посредством ее скрытой передачи и хранения внутри цифровых объектов различной природы. Помимо этого, подобные методы решают ряд вспомогательных задач, связанных с обработкой и обеспечением безопасности слабоструктурированной информации: встраивание в цифровые объекты метаданных различного формата, расследование случаев утечки конфиденциальной информации и выявление каналов утечки за счет встраивания в данные уникальных меток, ассоциированных с определенными пользователями.

4. Восстановление данных

Задача автоматической коррекции искаженных текстов [6] с незначительной долей случайных ошибок хорошо изучена, существуют готовые решения, использующиеся, например, при обработке поисковых запросов. Однако при высоких уровнях

искажений возможна неверная коррекция текста (в том числе, ошибочное изменение верных фрагментов искаженного текста), появление многозначности, когда принципиально разные варианты скорректированного текста могут хорошо согласовываться с используемыми моделями. Автоматический выбор истинного варианта становится затруднителен. При уровнях искажений, превышающих известную теоретико-информационную границу, среднее число возможных вариантов текста растет экспоненциально с ростом его длины, и однозначная коррекция невозможна.

Авторами предлагается разработка двух новых подходов, которые позволят получать информацию из зашумленного сообщения в условиях сильных шумов:

– фрагментарное восстановление зашумленного текста на тех участках, которые характеризуются низкой энтропией исходного текста;

– статистическое определение отдельных характеристик текста, в том числе его язык, жанр, авторство, стиль и др.

Обоснованием достижимости первого подхода может служить то, что локальные энтропийные характеристики текста являются нестационарными. Поэтому могут существовать отдельные низкоэнтропийные участки, для которых в силу теоретико-информационного неравенства возможно однозначное восстановление текста.

Обоснованием достижимости второго подхода может служить то, что такие характеристики текста как язык, жанр и др. могут быть идентифицированы статистическими способами, например, использующими распределение символьных N-грамм. Такие статистические методы являются устойчивыми к случайным искажениям.

В настоящее время уровень научной новизны определяется как лучшие мировые практики, так как существующие научные подходы ориентированы на использование экстенсивных решений – за счет использования значительных вычислительных ресурсов, в том числе более глубокого перебора и прохода деревьев в глубину, не учитывая окружающий контент и контекст.

5. Заключение

Авторами рассматривается задача обеспечения информационной безопасности в системах обработки слабоструктурированных данных. Перспективы развития таких систем диктуют актуальное направление и в область защиты данных. К особенностям упомянутых систем относятся объединение слабоструктурированных данных в единые базы с большими данными для комплексной обработки, нечеткое содержание данных, наличие искажений и необходимости восстановления. Системы информационной безопасности слабоструктурированных данных должны включать в себя решения, обеспечивающие как достаточный уровень защиты смоделированной инфраструктуры данных с учетом всех объединений и процессов, так и не препятствующих необходимым требованиям по скорости обработки и коммуникации системы с источниками и пользователем.

При этом слабоструктурированные данные за счет сложноформализуемых атрибутов могут позволить получать новые знания, предсказания трендов, что может помочь в принятии более обоснованных бизнес-решений, повышая эффективность работы.

Исходя из указанных аспектов, системы обработки слабоструктурированных данных будут продолжать развиваться и становиться более сложными и мощными, что откроет новые возможности для бизнеса и науки.

Работа выполнена при финансовой поддержке РНФ (проект № 24-11-00340)

Литература:

1. *Молокович О.А.* Подходы к извлечению информации из слабоструктурированных данных // Молодежный вестник УГАТУ. – 2021. – № 2 (25). – С. 64-66.

2. *Сомов С.К.* Репликация как инструмент повышения надежности функционирования распределенных систем // Информационные технологии и вычислительные системы. – 2018. – №3. – С. 69-79.

3. *Агаджанов А.О.* Проблемы автоматизированного сбора данных // Молодой ученый. – 2023. – № 49 (496). – С. 21-22.

4. *Менщиков А.А., Перфильев В.Э., Федосенко М.Ю., Фабзиев И.Р.* Основные проблемы использования больших данных в современных информационных системах // Столыпинский вестник. – 2022. – Т. 4. № 1. – URL: https://stolypin-vestnik.ru/wp-content/uploads/2022/01/31.pdf (дата обращения 16.09.2024).

5. *Evsutin O., Ivanov F., Dzhanashia K.* Watermarking for social networks images with improved robustness through polar codes // IEEE Access. – 2024. – Vol. 12. – P. 118154-118168.

6. *Заславская В.Л.* Проблемы использования слабоструктурированных и неструктурированных данных в системах бизнес-аналитики // Мягкие измерения и вычисления. – 2022. – Т. 59. № 10. – С. 84-94.

Домашкин А.Д., Логинова Л.Н.

Сравнительный анализ алгоритмов машинного обучения для обнаружения аномалий в информационных системах

Аннотация: В работе рассматривается сравнительный анализ алгоритмов машинного обучения для обнаружения аномалий в информационных системах, включая такие методы, как K-ближайших соседей, метод опорных векторов, кластеризация K-средних и изолированные деревья. Подчеркивается важность выбора подходящего алгоритма в зависимости от типа данных и требований к точности для повышения качества информационной безопасности.

Ключевые слова: информационные системы, машинное обучение, аномалии, информационная безопасность, задачи классификации

В условиях стремительного увеличения объема данных и сложности информационных систем (ИС) обеспечение их безопасности становится ключевым вопросом. Число кибератак только на российские ИТ-компании, которые используют современные методы защиты, во втором полугодии 2023 года выросло в 4 раза по сравнению с аналогичным периодом 2022-го и