УДК 004.056.55

Жидков А.А., бакалавр,
студен кафедры
«Робототехника и комплексная автоматизация»
Московский государственный технический
университет им. Н. Э. Баумана
Россия, г. Москва

Текстовая стеганография: обзор методов и перспективы развития

Аннотация

Статья представляет собой обзор современных и классических методов текстовой стеганографии — направления, связанного с сокрытием информации в тексте без изменения его смысла. Рассмотрены ключевые подходы: лингвистические (синонимия, вариации написания), синтаксические (манипуляции с пунктуацией и пробелами), методы на основе форматирования (шрифты, пробелы) и статистические алгоритмы. Особое внимание уделено их устойчивости к обнаружению, применимости в современных условиях и ограничениям. Обсуждаются перспективы развития текстовой стеганографии в контексте роста требований к информационной безопасности.

Ключевые слова: *текстовая стеганография, сокрытие данных, лингвистические техники, синтаксические методы, информационная безопасность.*

Text Steganography: A Review of Methods and Future Directions

Abstract

This article provides a comprehensive review of classical and modern text steganography methods, which focus on hiding information within text without altering its meaning. Key approaches are analyzed, including linguistic (synonym substitution, spelling variations), syntactic (punctuation and spacing manipulation), format-based (fonts, whitespace), and statistical techniques. The paper highlights

their resistance to detection, practical applicability, and limitations. Future directions for text steganography are discussed in the context of evolving information security demands.

Keywords: *text steganography, data hiding, linguistic techniques, syntactic methods, information security.*

Введение

Текстовая стеганография представляет собой метод скрытия информации, который может варьироваться от изменения форматирования существующего текста до модификации его содержания или создания нового текста. Для формирования естественно выглядящих текстов используются случайные последовательности символов или контекстно-свободные грамматики, что позволяет сохранять понятность исходного сообщения [1].

Особую сложность текстовой стеганографии создает отсутствие избыточной информации, характерной для мультимедийных файлов. В отличие от изображений или аудиозаписей, где внутренняя структура данных отличается от их визуального представления, текстовая информация отображается практически без изменений. Тем не менее, текстовые файлы обладают неоспоримыми преимуществами - они требуют меньше памяти для хранения и проще в передаче по сравнению с другими форматами данных.

Основные подходы в текстовой стеганографии включают лингвистические методы, основанные на особенностях языка, статистические алгоритмы генерации текста и техники работы с форматом представления информации. Каждый из этих подходов имеет свои характерные особенности и области применения, которые будут рассмотрены в данной статье. Механизм работы текстовой стеганографии представлен на рисунке 1.

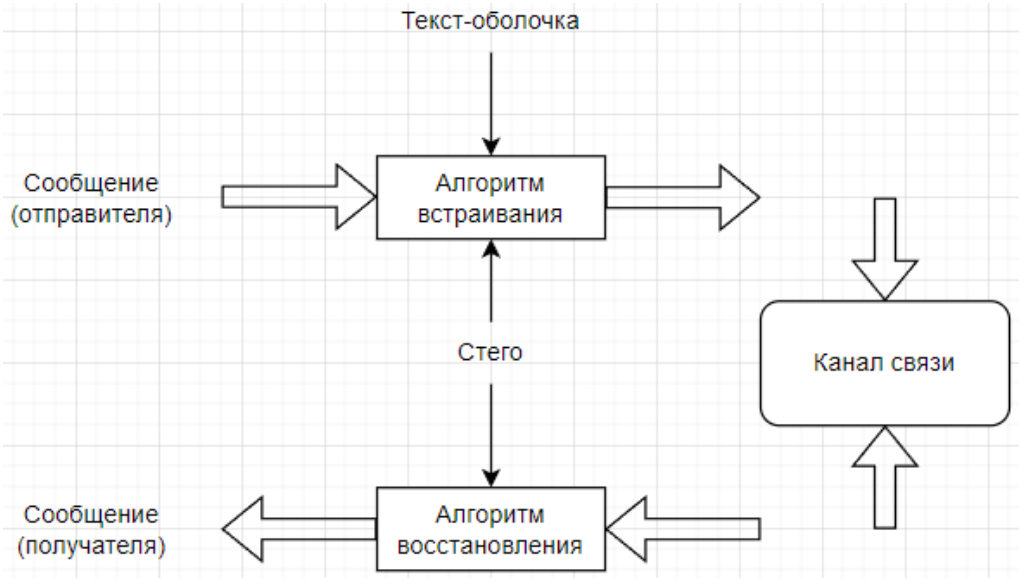

Рисунок 1 – Механизм текстовой стеганографии

Обзор методов

Метод вариативного правописания. Данный метод основан на использовании различий между американским и британским вариантами английского языка для сокрытия информации. Суть подхода заключается в замене слов на их альтернативные варианты написания, характерные для разных языковых стандартов. Например, слово "organize" (американский вариант) может быть заменено на "organise" (британский вариант), а "color" - на "colour" [2].

Эффективность метода обусловлена тем, что в определенных областях редко встречаются одновременно оба варианта написания. При этом для человека, не знакомого с особенностями обоих стандартов, подобные замены остаются незаметными. Однако данный подход имеет ограниченную надежность, так как различия в написании (например, добавление буквы "u" в британском варианте) могут быть легко выявлены при целенаправленном анализе.

Семантический метод. Данный подход основан на замене слов в тексте их синонимами для скрытия информации [2]. В отличие от метода вариативного правописания, где изменяется написание слов, здесь

используются разные слова с одинаковым или близким значением. Это позволяет сохранять смысл текста при внесении скрытых модификаций.

Ключевое преимущество метода - повышенная устойчивость к автоматическому обнаружению. Даже при использовании программ оптического распознавания символов (OCR) или других средств анализа, семантические замены остаются менее заметными по сравнению с изменениями в написании слов.

Однако эффективность метода напрямую зависит от правильного выбора синонимов. Неудачные замены, нарушающие естественность текста или его стилистику, могут привлечь внимание и привести к обнаружению скрытых данных. Особенно это актуально для специализированных текстов, где определенные термины и формулировки используются строго определенным образом.

Метод смещения строк. Данный подход основан на незначительном вертикальном смещении строк текста относительно их стандартного положения [2]. Каждая строка сдвигается вверх или вниз на строго определённую величину (например, 1/300 дюйма), что позволяет кодировать информацию за счёт микроскопических изменений в структуре документа. Для декодирования требуется точное измерение положения строк и выявление закономерностей их смещения.

Метод демонстрирует эффективность преимущественно при работе с печатными текстами, где исключено автоматическое преобразование средствами оптического распознавания (OCR). Однако его применение в цифровых документах ограничено, поскольку любые операции редактирования, повторное форматирование или конвертация файлов могут нарушить точность смещений и привести к потере скрытых данных.

Ключевой особенностью метода является сохранение визуальной целостности текста — изменения остаются незаметными при обычном просмотре. В то же время он уязвим к любым преобразованиям документа и

требует высокой точности оборудования для измерения позиций строк при восстановлении информации.

Метод сдвига слов. Данный подход основан на изменении горизонтальных промежутков между словами в тексте. Путем варьирования межсловных интервалов кодируется двоичная информация — увеличенное расстояние может соответствовать единице, а стандартное или уменьшенное — нулю. Особенность метода заключается в том, что естественная выравнивание текста при форматировании маскирует искусственно созданные промежутки [3], делая их незаметными для невооруженного глаза.

Метод наиболее эффективен в текстах с нефиксированным межсловным интервалом, где вариации расстояний не вызывают подозрений. Однако его устойчивость к обнаружению относительно невысока — при знании алгоритма кодирования или использовании специализированного ПО скрытые данные могут быть легко извлечены.

Ключевым ограничением является зависимость от неизменности форматирования. Любые операции редактирования, изменение шрифта или переформатирование текста приводят к нарушению заданных интервалов и потере скрытой информации. Кроме того, автоматическое выравнивание текста в современных текстовых редакторах может нивелировать искусственно созданные промежутки.

Синтаксический метод. Данный подход использует знаки препинания в качестве носителя скрытой информации. Путем стратегической замены точек, запятых и других символов пунктуации в тексте кодируются двоичные данные [3]. Например, точка может представлять ноль, а запятая — единицу, что позволяет незаметно встраивать сообщения в обычный текст без изменения его содержания.

Преимущество метода заключается в естественности внесенных изменений — вариации пунктуации редко вызывают подозрения при чтении. Особенно эффективно этот подход работает в художественных и публицистических текстах, где допустима некоторая свобода в использовании

знаков препинания. Однако его применение требует тщательного планирования, чтобы сохранить стилистическую целостность текста и избежать неестественного скопления или отсутствия пунктуационных знаков.

Основное ограничение метода связано с его чувствительностью к любым изменениям текста. Даже незначительное редактирование пунктуации — автоматическое или ручное — приводит к полной потере скрытых данных. Кроме того, информационная емкость метода относительно невелика, так как объем кодируемой информации напрямую зависит от количества знаков препинания в тексте.

Метод находит применение в ситуациях, когда текст передается без промежуточного редактирования, а его пунктуационная структура остается неизменной от отправителя к получателю. В таких условиях синтаксическое кодирование обеспечивает достаточную степень скрытности при минимальном вмешательстве в исходный текст.

Метод синонимической замены. Данный подход основан на замене слов в тексте их семантическими эквивалентами для скрытой передачи информации [4]. В отличие от методов, изменяющих написание слов, здесь используются полноценные синонимы, включая межъязыковые варианты. Например, американское "movie" может быть заменено британским "film", а "faculty" — эквивалентом "staff", что делает изменения менее заметными при чтении.

Основное преимущество метода заключается в повышенной скрытности — семантические замены сложнее обнаружить, чем орфографические вариации, поскольку они не нарушают визуального восприятия текста. Особенно эффективен метод в профессиональной и художественной литературе, где допустима вариативность лексики.

Однако метод требует значительных временных затрат на подбор подходящих синонимов, которые должны не только соответствовать скрываемому сообщению, но и органично вписываться в контекст. Неудачные

замены могут нарушить стилистическое единство текста или изменить его смысловые оттенки, что повышает риск обнаружения скрытого сообщения.

Применение метода ограничено необходимостью сохранения исходной структуры текста — любая редакторская правка или автоматическая обработка могут нивелировать внесенные изменения. Тем не менее, при работе с неизменяемыми документами метод синонимической замены демонстрирует более высокую надежность по сравнению с орфографическими техниками.

Механизмы сокрытия текстовой информации

Текстовая стеганография сталкивается с уникальной проблемой - необходимостью сохранять естественность текста при внедрении скрытых данных. Простейшие подходы, такие как акростихи или позиционное кодирование, оказываются крайне уязвимыми, поскольку сам принцип их работы становится слабым местом при известном алгоритме дешифровки.

Более совершенные техники опираются на малозаметные модификации текстовой структуры. Микроскопические изменения форматирования, включая регулировку межсимвольных интервалов и отступов, позволяют скрывать информацию, сохраняя визуальную целостность текста. Особую группу составляют методы, использующие общедоступные тексты-контейнеры, где секретные данные определяются через систему координат (страница, строка, позиция).

Отличительной чертой современных стеганографических подходов является их устойчивость к внешним воздействиям. Некоторые алгоритмы демонстрируют поразительную живучесть, сохраняя скрытые сообщения даже после многократной печати и сканирования документов. Однако эффективность любого метода остается напрямую зависимой от сохранения исходной структуры текста и точности соблюдения алгоритмов внедрения.

Методы на основе формата. Форматоориентированные подходы в стеганографии используют физические характеристики текста как носитель скрытой информации [4]. Эти методы работают с визуальным представлением

текста, сохраняя его семантическое содержание неизменным. Технически они реализуются через модификацию параметров форматирования - межсимвольных интервалов, размеров шрифтов, невидимых символов или преднамеренное введение специфических орфографических особенностей.

Особенность таких методов заключается в их двойственной природе обнаружения. В то время как человеческое восприятие часто пропускает незначительные изменения форматирования, автоматизированные системы анализа могут выявлять подобные модификации с высокой точностью. Это создает парадоксальную ситуацию, когда методы, наиболее устойчивые к визуальному обнаружению, оказываются наиболее уязвимыми перед машинной проверкой.

Эффективность форматных методов существенно зависит от типа документа и среды его передачи. Наибольшую результативность они демонстрируют в условиях, где исключены автоматизированная обработка и множественные преобразования файлов. При этом их ключевое преимущество - возможность внедрения в существующие тексты без необходимости создания специального контента-оболочки.

Случайные и статистические методы. Данный подход основан на генерации специальных текстов, предназначенных исключительно для сокрытия информации [5]. В отличие от методов, работающих с готовым контентом, здесь создается оригинальный текст, статистические характеристики которого соответствуют естественному языку. Это позволяет избежать проблем, связанных с сравнением модифицированного текста с его исходной версией.

Основная сложность метода заключается в необходимости точного воспроизведения языковых закономерностей. Генерируемый текст должен сохранять лексическое разнообразие, частотные характеристики слов и грамматическую структуру, типичные для обычных документов. При этом даже незначительные отклонения в статистике употребления языковых

единиц могут сделать текст подозрительным для автоматизированных систем анализа.

Важной особенностью этих методов является их зависимость от предварительной договоренности между участниками коммуникации. Сгенерированная последовательность символов лишь выглядит случайной для внешнего наблюдателя, тогда как для отправителя и получателя она содержит четкую структуру, позволяющую извлечь скрытое сообщение.

Эффективность метода существенно снижается при использовании шаблонных фраз или ограниченного набора лексики, что приводит к возникновению неестественных повторений. Оптимальные результаты достигаются при тщательном моделировании статистических параметров естественной речи и учете особенностей конкретного языкового регистра.

Кодирование текстовых признаков. Данный метод основан на модификации визуальных характеристик текста для скрытой передачи информации [6]. В отличие от содержательных или структурных подходов, здесь используются параметры отображения символов: гарнитура шрифта, его цвет, кернинг, межстрочные интервалы и другие типографские атрибуты.

Ключевое преимущество метода заключается в его устойчивости к визуальному обнаружению. Человеческое восприятие обычно не фиксирует незначительные изменения в оттенках цвета или микроскопические вариации размеров шрифта. При этом объем скрываемых данных может быть существенным за счет комбинирования различных параметров форматирования.

Однако метод демонстрирует критическую зависимость от сохранения исходного формата документа. Любые преобразования файла, включая перекодировку, печать с последующим сканированием или обработку системами оптического распознавания (OCR), приводят к необратимой потере скрытой информации. Особенно уязвимыми оказываются тонкие модификации, такие как вариации оттенков серого в черно-белом тексте или микроскопическое изменение межбуквенных интервалов.

Эффективность метода существенно повышается при работе с цифровыми документами в исходных форматах, исключающих промежуточные преобразования. Наибольшую практическую ценность подход представляет в условиях контролируемой документооборота, где гарантируется сохранение всех атрибутов форматирования на пути от отправителя к получателю.

Стеганография в языках разметки. Современные языки разметки предоставляют уникальные возможности для скрытой передачи информации благодаря особенностям их обработки. HTML как наиболее распространенный представитель этого класса обладает специфическими характеристиками, особенно полезными для стеганографических целей.

Основное преимущество HTML заключается в том, что он не различает регистр написания элементов разметки [7]. Например, варианты написания тега переноса строки (BR в верхнем, br в нижнем или Br в смешанном регистре) обрабатываются одинаково, что позволяет кодировать информацию в вариациях их написания. Дополнительные возможности предоставляют теги форматирования (жирный шрифт, подчеркивание, курсив), которые могут быть добавлены без видимого изменения текста, но при этом нести скрытую информацию.

В отличие от HTML, XML как язык, чувствительный к регистру, менее подходит для стеганографии, поскольку требует строгого соблюдения синтаксиса. Именно гибкость HTML в этом отношении делает его основным инструментом - вариативность написания элементов сохраняет функциональность документа, одновременно позволяя внедрять скрытые сообщения.

При практическом применении метода необходимо учитывать несколько факторов: сохранение корректности конечного документа, обеспечение естественного распределения модифицированных элементов разметки и учет особенностей обработки разметки различными программами.

Основной риск связан с возможной автоматической стандартизацией кода системами управления контентом или валидаторами, которые могут автоматически приводить регистр элементов к единому стандарту. Поэтому метод наиболее эффективен в средах, гарантирующих сохранение исходной структуры HTML-документа.

Метод символических последовательностей. Данный подход основан на искусственном создании текстовых последовательностей, имитирующих статистические характеристики естественного языка. В отличие от методов, работающих с готовым текстом, здесь конструируются специальные символьные комбинации, сохраняющие типичные языковые закономерности - частотность букв, среднюю длину слов и другие лингвистические параметры.

Особенность метода заключается в его способности генерировать правдоподобный текст-оболочку без смыслового содержания [8]. Создаваемые последовательности символов формально соответствуют словарному составу языка, сохраняя при этом заданные стеганографические параметры. Например, могут искусственно воспроизводиться характерные для конкретного языка сочетания букв и распределение длин слов.

Ключевое преимущество - высокая устойчивость к статистическому анализу, поскольку генерируемый текст повторяет основные количественные характеристики естественной речи. Однако метод требует тщательной настройки генератора символов и глубокого знания лингвистической статистики целевого языка. Малейшие отклонения в частотности символов или структуре слов могут сделать текст подозрительным для автоматизированных систем проверки.

Основные сложности применения связаны с:

- Необходимостью точного воспроизведения языковых статистик
- Риском создания неестественных сочетаний при ручном конструировании
- Зависимостью от предварительных договоренностей о способе кодирования

Метод находит применение в условиях, требующих генерации больших объемов текста-контейнера с гарантированными лингвистическими характеристиками. Его эффективность повышается при использовании специализированного программного обеспечения, способного точно моделировать языковые закономерности.

Лингвистические методы стеганографии. Данный класс методов использует естественные свойства языка для скрытой передачи информации [9]. Основной принцип заключается в кодировании данных через манипуляции с лексическими единицами, где каждое слово или его характеристики могут нести определенные биты информации. Для этого применяются специальные кодовые книги, устанавливающие соответствие между языковыми элементами и битовыми последовательностями.

Ключевая особенность лингвистических подходов - их опора на естественные языковые структуры. В отличие от форматных методов, они работают непосредственно с содержательной частью текста, что теоретически должно обеспечивать большую скрытность. Однако на практике возникает парадоксальная ситуация: чем точнее метод воспроизводит языковые закономерности, тем сложнее обеспечить достаточную информационную емкость.

Семантический вариант метода, основанный на синонимических заменах, представляет более совершенный подход. Замена предопределенных слов их синонимами позволяет сохранять исходный смысл текста при внедрении дополнительной информации. Например, чередование слов "автомобиль" и "машина" в строго определенных позициях может кодировать бинарную последовательность.

Основные проблемы лингвистической стеганографии связаны с:

- Необходимостью сохранения смысловой целостности текста
- Ограниченным набором слов, пригодных для замены
- Риском нарушения стилистического единства
- Сложностью автоматизации процесса кодирования/декодирования

Эффективность методов существенно зависит от типа текста и богатства его лексики. Наибольшие результаты достигаются в специализированных областях с развитой синонимией и терминологической вариативностью. При этом остается актуальной проблема баланса между скрытностью, емкостью и устойчивостью к анализу.

Заключение

Текстовая стеганография занимает особое положение среди методов скрытия информации благодаря своей универсальности, доступности и минимальным требованиям к вычислительным ресурсам. В отличие от мультимедийных носителей, текстовые данные не обладают выраженной избыточностью, что делает задачу скрытия информации в них более сложной и одновременно более изящной с точки зрения реализации.

Проведённый в статье обзор продемонстрировал, что текстовая стеганография может опираться на широкий спектр подходов — от орфографических вариаций и семантических замен до форматных и синтаксических модификаций. Каждый метод обладает своими достоинствами и ограничениями, связанными с объемом скрываемой информации, устойчивостью к автоматическим преобразованиям, заметностью для читателя и степенью сложности реализации. Наиболее простые подходы, такие как акростихи и смещение строк, характеризуются низкой устойчивостью к внешним воздействиям, в то время как сложные семантические и статистические методы демонстрируют высокую скрытность, но требуют значительных интеллектуальных и вычислительных усилий.

Современное развитие текстовой стеганографии связано с необходимостью учета множества факторов — от стилистической целостности текста до устойчивости скрытых данных при передаче и возможном редактировании. Особенно важным направлением становится повышение живучести встроенной информации при сохранении

естественности текста, а также противостояние автоматизированным системам анализа, включая машинное обучение и лингвистические модели.

Несмотря на прогресс, необходимо признать, что текстовая стеганография остаётся сравнительно уязвимой областью. Автоматическая нормализация текстов, работа алгоритмов проверки грамматики и орфографии, а также стандарты форматирования в современных редакторах могут невольно разрушить тщательно встроенное сообщение. В связи с этим особую актуальность приобретают методы, устойчивые к преобразованиям и не зависящие от конкретной программной среды.

Таким образом, будущее текстовой стеганографии, вероятнее всего, связано с гибридными методами, сочетающими форматные и семантические подходы, а также с развитием систем, способных адаптироваться к изменениям в структуре текста. Интеграция методов машинного обучения для автоматической генерации стеганографически устойчивых текстов также представляется перспективной. В этом контексте текстовая стеганография сохраняет потенциал как инструмент обеспечения конфиденциальности, защиты авторских прав и безопасной передачи информации в условиях открытых каналов связи.

Использованные источники:

1. Fridrich, J. (2009). Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge University Press
2. Chapman, M. & Davida, G. (2002). Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text.
3. Taskiran, C. (2006). Text Steganography with Natural Language Processing.
4. Huang, D. et al. (2011). Text Steganography Through Changing Font Color. IEEE.
5. Low, S.H. (2001). Document Identification for Copyright Protection Using Centroid Detection.
6. Kim, Y. (2003). A New Text Steganography Using Punctuation Marks.

7. Bolshakov, I.A. (2004). A Method of Linguistic Steganography Based on WordNet.
8. Wayner, P. (2009). Disappearing Cryptography.
9. Alotaibi, R. & Elrefaei, L. (2019). Improved Capacity Arabic Text Steganography.

Email: zhidkovaa@student.bmstu.ru