**Evaluation Only. Created with Aspose.Words. Copyright 2003-2018 Aspose Pty Ltd.**

**Evaluation Only. Created with Aspose.Words. Copyright 2003-2018 Aspose Pty Ltd.**

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ

ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра информационных систем и технологий

Реферат

по дисциплине «Основы информационных технологий»

на тему «Анализ и сравнительная характеристика методов и инструментальных средств текстовой стеганографии»

Выполнил: магистрант факультета ИТ

специальности «Теоретические основы информатики»

Берников В.О.

Проверил: профессор,

доктор технических наук

Урбанович П.П.

Минск 2017

# **Введение**

Задача защиты информации от несанкционированного доступа решалась во все времена на протяжении истории человечества. Уже в древнем мире выделилось два основных направления решения этой задачи, существующие и по сегодняшний день: криптография и стеганография. Целью криптографии является скрытие содержимого сообщений за счет их шифрования. В отличие от этого, при стеганографии скрывается сам факт существования тайного сообщения [1].

Можно выделить две причины популярности исследований в области стеганографии в настоящее время: ограничение на использование криптосредств в ряде стран мира и появление проблемы защиты прав собственности на информацию, представленную в цифровом виде. Первая причина повлекла за собой большое количество исследований в духе классической стеганографии (то есть скрытия факта передачи информации), вторая – еще более многочисленные работы в области так называемых водяных знаков. Цифровой водяной знак (ЦВЗ) – специальная метка, незаметно внедряемая в изображение или другой сигнал с целью тем или иным образом контролировать его использование.

Принимая во внимание вышесказанное, можно сделать вывод о том, что на сегодняшний день актуальна научно-техническая проблема усовершенствования алгоритмов и методов проведения стеганографического скрытия конфиденциальных данных или защиты авторских прав на определенную информацию.

Целью реферата является анализ и сравнительная характеристика методов и инструментальных средств текстовой стеганографии.

Для достижения поставленной цели необходимо решить ряд задач:

* охарактеризовать основные методы текстовой стеганографии, существующие на данный момент;
* найти и проанализировать существующие программные средства компьютерной стеганографии, позволяющие скрывать информацию;
* выявить основные достоинства и недостатки существующих программных средств.

# **Текстовая стеганография. Классификация и общая характеристика методов**

Стеганография – это наука о способах передачи (хранения) сокрытой информации, при которых скрытый канал организуется на базе и внутри открытого канала с использованием особенностей восприятияинформации, причем для этой цели могут использоваться такие приемы, как:

* полное сокрытие факта существования скрытого канала связи;
* создание трудностей для обнаружения, извлечения или модификации передаваемых сокрытых сообщений внутри открытых сообщений-контейнеров;
* маскировки сокрытой информации в протоколе.

Компьютерная стеганография изучает способы сокрытия информации в компьютерных данных, представляющих собой различные файлы, программы, пакеты протоколов и т. п.

Современная стеганография, как правило, имеет дело с электронными средствами. Компьютерная стеганография базируется на двух принципах. Первый заключается в том, что файлы, содержащие оцифрованное изображение или звук, могут быть до некоторой степени видоизменены без потери функциональности, в отличие от других типов данных, требующих абсолютной точности.

Второй принцип состоит в неспособности органов чувств человека различить незначительные изменения в цвете изображения или качестве звука, что особенно легко использовать применительно к объекту, несущему избыточную информацию, будь то 16-битный звук, 8-битное или, еще лучше, 24-битное изображение. Если речь идет об изображении, то изменение значений наименее важных битов, отвечающих за цвет пиксела, не приводит к сколь-нибудь заметному для человека изменению цвета.

Особенностью стеганографического подхода является то, что он не предусматривает прямого оглашения факта существования защищаемой информации. Это обстоятельство позволяет в рамках традиционно существующих информационных потоков или информационной среды решать некоторые важные задачи защиты информации ряда прикладных областей.

Основным определяющим моментом в стеганографии является стега-нографическоепреобразование. До недавнего времени стеганография, как наука, в основном изучала отдельные методы сокрытия информации и способы их технической реализации. Разнообразие принципов, заложенных в стеганографических методах, по существу тормозило развитие стеганографии как отдельной научной дисциплины и не позволило ей сформироваться в виде некоторой науки со своими теоретическими положениями и единой концептуальной системой, которая обеспечила бы формальное получение качественных и количественных оценок стеганометодов. В этом история развития стеганографии резко отличается от развития криптографии.

Стеганографическая система (стегосистема) – совокупность средств и методов, которые используются для формирования скрытого канала передачи информации.

Стегосистема образует стегоканал, по которому передается (или в котором хранится) заполненный контейнер. Этот канал считается подверженным воздействиям со стороны нарушителей.

При построении стегосистемы должны учитываться следующие положения:

* противник имеет полное представление о стеганографической системе и деталях ее реализации, единственной информацией, которая остается неизвестной потенциальному противнику, является ключ, с помощью которого только его держатель может установить факт присутствия и содержание скрытого сообщения;
* если противник каким-то образом узнает о факте существования скрытого сообщения, это не должно позволить ему извлечь подобные сообщения в других данных до тех пор, пока ключ хранится в тайне;
* потенциальный противник должен быть лишен каких-либо технических и иных преимуществ в распознавании или раскрытии содержания тайных сообщений.

Любая стегосистема должна отвечать следующим требованиям:

* свойства контейнера должны быть модифицированы, чтобы изменение невозможно было выявить при визуальном контроле. Это требование определяет качество сокрытия внедряемого сообщения: для обеспечения беспрепятственного прохождения стегосообщения по каналу связи оно никоим образом не должно привлечь внимание атакующего;
* стегосообщениедолжно быть устойчиво к искажениям, в том числе и злонамеренным. В процессе передачи изображение (звук или другой контейнер) может претерпевать различные трансформации: уменьшаться или увеличиваться, преобразовываться в другой формат и т. д. Кроме того, оно может быть сжато, в том числе и с использованием алгоритмов сжатия с потерей данных;
* для сохранения целостности встраиваемого сообщениянеобходимо использование кода с исправлением ошибки;
* для повышения надежности встраиваемое сообщение должно быть продублировано.

Современная стеганография, как правило, «имеет дело» с электронными средствами. Это может объясняться следующими причинами. Во-первых, так как объем осаждаемой информации, как правило, довольно небольшой по сравнению с размером контейнера, в котором она будет скрыта, то в электронные контейнеры гораздо проще скрывать данные и извлекать их. Во-вторых, процедура осаждения/извлечения может быть автоматизирована с помощью специальных программных средств. В-третьих, электронный формат данных характеризуется информационной избыточностью, которой можно управлять, чтобы скрыть сообщения. Эти общие для всех электронных документов особенности ставят знак равенства между цифровой и компьютерной стеганографией.

На рисунке 1.1 приведена наиболее общая классификация методов. Метка содержит скрытую информацию, подтверждающую авторство. Скрытая информация призвана обеспечить защиту прав интеллектуальной собственности [2].

В качестве внедряемой информации можно использовать данные об авторе, дату и место создания произведения, номера документов, подтверждающих авторство. В качестве иллюстрации более подробно рассмотрим использование стеганографии для защиты авторских прав на текстовые документы.



Рисунок 1.1 ­­­– Общая классификация методов компьютерной стеганографии

Компьютерная стеганография базируется на двух принципах. Первый заключается в том, что если файлы, содержащие оцифрованное изображение или звук, могут быть до некоторой степени видоизменены без потери функциональности, то текстовые документы, коды программ или базы данных требуют абсолютной точности при обратных преобразованиях. Это обстоятельство чрезвычайно важно, если, например, текстовый документ-контейнер с осажденной информацией претерпевает конвертацию на основе иного стиля (шрифта, кегля и т.п.) или при его архивации. Понятно, что сам документ при этом не должен измениться.

Все многообразие методов текстовой стеганографии подразделяется на синтаксические, которые не затрагивают семантику текстового сообщения, и лингвистическиеметоды, которые основаны на эквивалентной трансформации текстовых файлов, сохраняющей смысловое содержание текста, его семантику.

## **Синтаксические методы текстовой стеганографии**

К числу этих методов этого подкласса относят следующие.

**Line-shift coding**(изменение расстояния между строками электронного текста). Он также называется методом изменения межстрочных интервалов. Его сущность заключается в том, что используется текст с различными межстрочными расстояниями. Выделяется максимальное и минимальное расстояния между строками, позволяющее кодировать соответственно символы «1» и «0» осаждаемого сообщения (рисунок 1.2). Разница в межстрочных расстояниях авторами изменялась на 1/300 дюйма (это расстояние было привязано к существовавшей в то время разрешающей способности монитора) [3].

Очевидным недостатком метода является его низкая эффективность: размер в битах осаждаемой информации не может превысить количество строк в контейнере.



Рисунок 1.2 – Иллюстрация метода на основе изменения межстрочных интервалов

**Word-shift coding**(изменение расстояния между словами в одной строке электронного текста). Суть метода состоит в том, что осаждение информации основано на модификации расстояния между словами текста-контейнера. Аналогично предыдущему методу, выделяется максимальное и минимальное расстояния между словами, обозначающие соответственно символ «1» и «0», и остальные расстояния, или некоторые из них, увеличивают или уменьшают до размеров уже выделенных. Частный случай этого метода – метод изменения количества пробелов(рисунок 1.3). Данный рисунок показывает пример внедрения в текст-контейнер бинарной последовательности 0101100100111010. Как видно, переход с одинарного пробела на двойной кодирует «1» (пары выделены красным фоном), переход же с двойного пробела на одинарный кодирует «0» (пары выделены синим фоном).





Рисунок 1.3 – Пример использования метода на основе длин пробелов

Данный метод имеет недостатки. Во-первых, он мало эффективен, так как необходим контейнер большого объема (объем скрытых данных в данном случае приблизительно равен одному биту на 160 байт текста). Во-вторых, возможность сокрытия зависит от структуры текста (некоторые тексты, например, белые стихи, не имеют четких признаков конца). В-третьих, текстовые редакторы часто автоматически добавляют символы пробела после точки.

**Feature coding**(внесение специфических изменений в шрифты (начертания отдельных букв). Этот метод заключается в изменении написания отдельных букв используемого стандартного шрифта (рисунок 1.4).





Рисунок 1.4 – Иллюстрация метода feature coding

Визуально заметны различные образы, соответствующие буквам с верхними (напр., *l*, *t*, *d*) или нижними (напр., *a*, *g*) выносными элементами. Например, букву «А» можно модифицировать, незначительно укорачивая длинную нижнюю часть буквы. При этом можно закодировать стегосообщение так, что модифицированная буква будет означать «1», а немодифицированная — «0».

Модифицировать можно несколько букв. Таким образом, объем встраиваемого сообщения будет увеличиваться.

Результат внедрения секретного сообщения «1» в текст-контейнер «А», при использовании метода feature coding и текстового процессора MS Office Word 2007, показан на рисунке 1.5 (а) пустой контейнер; б) заполненный контейнер (со стегосообщением «1»)).



Рисунок 1.5 – Пример применения метода feature coding

**Метод изменения интервала табуляции** аналогичен вышеописанному методу изменения количества пробелов, только в этом случае меняется не количество пробелов, а соответственно расстояние между строками и интервал табуляции [4].

**Null chipper**(дословно – несуществующий, нулевой лепет). Предполагает размещение тайной информации на установленных позициях слов или в определенных словах текста-контейнера, который, как правило, лишен логического смысла (как видно, действительно лепет). На рисунке 1.6 показан пример реализации метода — скрытой информацией являются первые символы слов.



Рисунок 1.6 – Пример реализации метода null chipper

**Метод увеличения длины строки**. Предусматривает искусственное увеличение длины каждой строки за счет пробелов: например, одному пробелу соответствует логический 0, двум — 1 (рисунок 1.7).

Преимущество такого метода кодирования состоит в том, что оно может быть выполнено с любым текстом, изменения в формате резко не бросаются в глаза читателю, обеспечивается передача большего числа скрытых данных по сравнению с предыдущим методом (примерно 1 бит на 80 байт содержимого контейнера).



Рисунок 1.7 – Иллюстрация метода увеличения длины строки

Недостаток метода состоит в том, что некоторые компьютерные программы (например, Sendmail) могут неосторожно удалять дополнительные пробелы. Помимо этого, скрытые таким образом данные не всегда могут быть восстановлены с печатной копии документа.

**Использование регистра букв**. Для обозначения бита секретного сообщения, представленного единицей, используется символ нижнего регистра, а нулем — верхнего (или наоборот). Например, секретный текст, состоящий из одной буквы «А», необходимо внедрить в текст-контейнер «steganography». Для этого используется двоичное представление кода символа «А» — «01000001». Далее предположим, что для обозначения бита секретного сообщения, представленного единицей, используется символ верхнего регистра, а нулем — нижнего.

Результат внедрения секретного сообщения «А» в текст-контейнер «steganography» показан на рисунке 1.8.



Рисунок 1.8 – Иллюстрация метода на основе регистра клавиатуры

Существует модификация данного метода, основанная на применении различных алфавитов, в которых используются символы, имеющие одинаковое начертание, но различную кодировку (например, *а*, *в*, *е*, *р*, *т*, с).

**Метод невидимых символов**. Пробел кодируется символом с кодом 32, но в тексте его можно заменить также символом, имеющим код 255 (или 0), который является «невидимым» и отображается как пробел.

Методы могут применяться независимо и совместно, сохраняют исходный смысл текста, а обеспечиваемые ими показатели плотности кодирования при совмещении складываются.

Рассмотренные методы работают успешно до тех пор, пока тексты представлены в коде ASCII. Существуют также стеганографические методы, которые интерпретируют текст как двоичное изображение. Необходимо отметить, что данные методы нечувствительны к изменению масштаба документа, что обеспечивает им хорошую устойчивость к большинству искажений, которые могут иметь место при активных атаках.

Описанные выше методы – синтаксические – легко применяются к любому тексту, независимо от его содержания, назначения и языка. Синтаксические системы стеганографии легко реализуются в программном коде, так как они полностью автоматические и не требуют вмешательства оператора. Однако синтаксические методы неустойчивы к форматированию текста (вспомним системы на основе ЦВЗ), и поэтому информация может быть потеряна при простом применении иного стиля форматирования текста-контейнера, скрывающего в себе стегосообщение. К тому же с помощью синтаксических методов можно передать незначительное количество информации. В литературе отсутствуют результаты экспериментального исследования эффективности проанализированных методов с проверкой на больших объемах данных [5].

## **Лингвистические методы текстовой стеганографии**

Стеганографические методы, основанные на лексической структуре текста, обладают большими возможностями.

Лингвистическая стеганография занимается скрытым кодированием произвольной информации, представленной в двоичном виде, в текстах.

Необходимо отметить, что осмысленность и внешняя «безобидность» текста должна сохраниться. К лексическимметодам встраивания скрытой информации в текстовые файлы-контейнеры относят метод переменной длины слова, метод первой буквы, метод синонимов и другие.

Одним из наиболее обсуждаемых методов является метод, основанный на системе синонимов языка, используемого для написания электронного текста. Проведенные исследования для случая английского языка, показали, что среднее количество синонимов в одном подмножестве синонимов равняется 2,56. Минимальное количество синонимов в одном множестве синонимов равняется 2, а максимальное 13.

В качестве примера приведем множество синонимов S0: {«propensity», «predilection», «penchant», «proximity»}. В приведенном множестве синонимов, каждое слово имеет единственное одинаковое смысловое значение, что позволяет закодировать каждое слово своим уникальным кодом, например, «propensity» - 00, «predilection» - 01, «penchant» - 10, «proximity» - 11. Подобное кодирование позволяет выбирать одно из четырех слов в зависимости от двух бит секретного сообщения. Отметим, что при этом, независимо какое из четырех слов будет выбрано, семантика (смысл) сообщения не изменится.

Отправитель и получатель имеют одинаковое множество синонимов, поддерживаемое одним и тем же электронным словарем.

Процедура передачи секретного сообщения, используя лексическую стеганографию, производится в следующей последовательности:

1. Первоначально отправитель выбирает контейнер (текстовый файл).

2. Отправитель преобразует секретное сообщение в двоичную последо-вательность: ...01000... , используя по возможности криптографические методы.

3. Отправитель, последовательно анализируя текстовый файл, находит первое слово, для которого существуют *N* синонимов.

4. Отправитель вычисляет целую часть значения *log2 N,* которая определяет число символов секретного сообщения, которые могут быть внедрены в контейнер путем выбора соответствующего синонима. Например, если в тексте встретилось слово «penchant», принадлежащее множеству *S0:* {«propensity», «predilection», «penchant», «proximity»}, в соответствии со значением двух бит закодированной секретной информации - 01, это слово должно быть заменено словом синонимом «predilection» [6].

Аналогичные действия выполняет и получатель. Получатель анализирует слова в контейнере (текстовом файле) на предмет принадлежности к множеству синонимов. Если текущее слово относится к одному из множеств синонимов, он определяет мощность этого множества *N.* Целая часть *log2N* определяет число бит, которые закодированы на основании текущего множества синонимов. Например, если получатель обнаружит в тексте слово "predilection" и определит, что оно относится к множеству синонимов *S0* состоящему из *N*=4 синонимов, тогда *log2 N=2* бита, а слово "predilection" интерпретируется, как два бита (01) секретного сообщения. Следует отметить, что в каждом подмножестве синонимов их упорядочивание должно выполняться по одному и тому же алгоритму и у отправителя сообщения, и у его получателя. Например, в алфавитном порядке [7].

В случае слов с несколькими смысловыми значениями подобное кодирование оказывается невозможным. Также невозможно кодирование, если один из синонимов состоит из двух (или более), разделенных пробелом, слов.

К сожалению, количество синонимов не всегда равно 2k, где *к* - целое положительное число. Например, имеем три подмножества синонимов S0: {AAA, ВВВ, ССС}; S1{МММ, NNN, ООО, РРР, QQQ}; S2:{WWW, XXX, YYY}.

Использование данных подмножеств синонимов согласно приведенной ранее идеи семантической стеганографии позволяют закодировать один бит на основании первого и третьего подмножеств и два бита за счет второго подмножества. Всего может быть закодировано только четыре бита, в то время как суммарная мощность приведенных подмножеств позволяет закодировать 3\*5\*3=45 состояний или не менее, чем |log2 45|=5 бит.

Для увеличения объема внедряемой в контейнер секретной информации может быть использована система счисления со смешанным основанием. В этом случае работают не с одним множеством синонимов, а с группой множеств синонимов. Кодирование выполняется не путем изолированного использования каждого множества синонимов, а используя их совокупности. Например, для приведенных трех множеств можно закодировать log2 45 = 5 бит (берется целочисленное значение) вместо 4. Для этого используется код, состоящий из трех цифр, в котором первая и последняя цифры могут принимать значения от 0 до 2, а средняя – от 0 до 4.

Отличительной особенностью методов лексической стеганографии является то, что пользователь, как правило, должен сам составлять (или видоизменять) тот объект (текст-контейнер), в котором будет передаваться или храниться тайная информация.

Так, при использовании **метода переменной длины**пользователю, который хочет послать секретное сообщение, необходимо сгенерировать (набрать) текст, в котором слова должны иметь соответствующую длину. Длина этих слов зависит от секретного сообщения и способа кодирования. Обычно одно слово текста-контейнера определенной длины кодирует два бита информации из стегосообщения. Например, слова текста длиной в 4 и 8 символов могут означать комбинацию бит «00», длиной в 5 и 9 – «01», 6 и 10 – «10», 7 и 11 букв – «11». Слова короче 4 и длиннее 11 букв можно вставлять где угодно для лексической и грамматической связки слов в предложении. Программное приложение, которое декодирует принятое сообщение, будет просто игнорировать их.

При использовании **метода первой буквы**можно передавать еще больше скрытой информации в одном слове: обычно это три или четыре бита. Программа-помощник в этом методе накладывает ограничение уже не на длину слова, а на первую (можно на вторую) букву. Обычно одну и ту же комбинацию могут кодировать несколько букв, например, комбинацию «101» означают слова, начинающиеся с «А», «Г» или «Т». Это дает большую свободу выбора оператору, придумывающему стегосообщение, и текст не будет нелепым, не содержащим смысла.

Другим, не менее распространенным лексическим методом передачи скрытой информации является мимикрия***.*** Мимикрия генерирует осмысленный текст, используя синтаксис, описанный в **Context Free Grammar**(CFG), и встраивает информацию, выбирая из CFG определенные фразы и слова. Грамматика CFG – это один из способов описания языка, который состоит из статических слов и фраз языка, а также узлов. Узлы в простейшем случае представляют собой места в генерируемом тексте, где может быть принято решение, какое слово или фразу дальше необходимо вставить в текст. Мимикрия создает бинарное дерево, которое основано на возможностях CFG и составляет текст, выбирая те листья дерева, которые кодируют нужный бит.

Например, необходимо передать секретное сообщение 10101, используя следующее бинарное дерево:

Старт -> существительное

существительное -> Илья || Иван

глагол -> поехал куда || пошел куда

куда -> на работу, чтобы зачем. || домой, чтобы зачем.

зачем -> забрать что || взять что

что -> деньги || одежду.

Основываясь на приведенном бинарном дереве, покажем процедуру внедрения секретного кода «10101», приведенной в таблице 1.1.

Таблица 1.1 – Внедрение секретного кода на основе бинарного дерева



Окончательно получилось следующее предложение: *Иван поехал домой, чтобы забрать одежду*.

Недостатками этого метода является то, что с его помощью нельзя передавать большие объемы информации, а также низкая производительность метода. Кроме того, необходимо отметить невысокую скрытность секретного сообщения, которое в сильной мере влияет на структуру передаваемого текста.

Еще одним методом, основанным на мимикрии, является **Spammimic**(уподобление спаму) (рисунок 1.9).



Рисунок 1.9 – Текст-контейнер с осажденным в нем сообщением

«*Здравствуйте*!»

Здесь в качестве контейнера используется обычный спам внутри которого размещаются установленным обеими сторонами способом значащие символы (стегосообщение).

Существует и множество других методов преобразования текста. В любом случае, при разработке эффективных лексических стеганографических методов, необходимо искать золотую середину: контейнер должен быть «плотно» насыщен стегоинформацией и при этом совершенно не должен выделяться из обычной общей массы файлов такого же формата и наполнения.

В таблице 1.2 представлены результаты сравнительного анализа эффективности некоторых из проанализированных методов.

Таблица 1.2 – Сравнительные результаты анализа эффективности синтаксических методов текстовой стеганографии

|  |  |  |  |
| --- | --- | --- | --- |
| Метод | Стегознаков | Плотность заполнения, % | Часть стегосообщения, содержащаяся в каждом символе контейнера, бит |
| Line-shift coding | 811998 | 1,2 | 0,013 |
| Word-shift coding | 8349980 | 13,1 | 0,132 |
| Feature coding | 49145754 | 77,6 | 0,776 |
| Метод изменения регистра буквы | 49145754 | 77,6 | 0,776 |
| Метод изменения цвета символов | 63328767 | 100,0 | 15,000 |
| Метод изменения масштабов символов | 63328767 | 100,0 | 3,000 |
| Изменение цветовых координат | 63328767 | 100,0 | 3,000 |

Под «плотностью заполнения» (третий столбец таблицы) понимается отношение стегознаков в пустом контейнере к общему числу символов в пустом контейнере. Например, для метода изменения регистра буквы и метода feature coding стегознаками являются буквы любого алфавита, т.е. пробелы, цифры, специальные знаки и символы не учитываются. А в методах изменения цвета символов, изменения масштаба символов стегознаками являются все символы документа, в том числе специальные знаки и символы.

До сих пор вопрос о создании безопасной лингвистической стегосистемы остается открытым. Любая обработка текста редактором, его печать или перевод в другой формат может изменить расположение пробелов и уничтожить скрытый текст.

**This document was truncated here because it was created in the Evaluation Mode.**