УДК 003.26

Применение статистического анализа для обнаружения скрытых сообщений в текстовых данных [1]

И. В. Нечта

В данной статье предлагается улучшение предложенного ранее метода стегоанализа текстовых данных. Суть предыдущего метода состояла в том, что извлекаемое из текста сообщение проверялось на случайность при помощи статистического теста хи-квадрат. Случайность извлечённого сообщения означала наличие внедрения. В настоящей работе предлагается модифицировать тест хи-квадрат таким образом, чтобы используемый в тесте квантиль был подобран эмпирическим путём. Такая модификация уменьшает вероятность ошибки стегоанализа.

Ключевые слова: Tyrannosaurus Lex, стеганография, стегоанализ, стеготекст.

1. Введение

На сегодняшний день для обеспечения информационной безопасности систем используют различные средства и методы. Среди них особую роль занимают методы стеганографии. В отличие от криптографии, ограничивающей доступ к информации, содержащейся в передаваемом сообщении с помощью некоторого секретного ключа, задача стеганографии состоит в том, чтобы скрыть сам факт передачи какого-либо сообщения от третьих лиц. Обычно такая задача решается путём внедрения передаваемого секретного сообщения в безобидный на вид объект данных, так называемый *контейнер*. Сам контейнер подбирается таким образом, чтобы факт его существования или передача не вызывали никакого подозрения. Основными характеристиками методов стеганографии следует считать объём внедряемого сообщения и устойчивость к анализу – обнаружению факта наличия внедрения.

Одним из активно развивающихся направлений стеганографии является цифровая стеганография. В этом направлении в качестве контейнера используется цифровой объект – компьютерный файл. Современные методы встраивания позволяют внедрять информацию в файлы звука, видео, текста, исполняемых программ и т.д. Основной принцип работы алгоритмов внедрения скрытой информации в цифровые объекты базируются на возможности незначительного изменения файла так, чтобы его содержимое казалось неизменным с точки зрения внешнего наблюдателя. Например, при встраивании информации в изображение используются методы замены младших значащих бит (LSB-внедрение). На современном этапе развития LSB-методов считается, что встраивание должно осуществляться не поочерёдно в каждый пиксель, а в случайном порядке. Также накладываются существенные ограничения на объём внедряемой информации. При внедрении сообщений в видеофайлы используют встраивание дополнительных кадров. Кроме того, может быть применён следующий подход. Съёмка фильма ведётся сразу же с нескольких камер, которые стоят рядом друг с другом. На этапе монтажа готовый фильм собирается из набора отснятых фрагментов. Выбор камеры для соответствующего фрагмента определяется внедряемым сообщением. Очевидно, что человеческий глаз неспособен уловить столь малые изменения позиций

[1] Работа выполнена в рамках НИР по госконтракту № 02.740.11.0396 и гранта РФФИ № 09-07-00005.

съёмки сцены. Для внедрения информации в текстовые данные широко используется, например, метод замены слов на их синонимы. Сначала в текстовом контейнере производится поиск слов, которые имеют синонимы. Затем, в соответствии со скрываемым сообщением, производится замена найденных слов на соответствующие им синонимы. Стоит отметить, что смысловое содержание текста остается неизменным, делая метод встраивания достаточно устойчивым к обнаружению. Исполняемые программы также можно использовать в качестве контейнеров. Однако, в отличие от видеофайлов и изображений, исполняемые файлы имеют свою специфику. Незначительное изменение исполняемого файла может разрушить алгоритм его работы и сделать программу неработоспособной, что является недопустимым. Существующие подходы встраивания используют замену вычислительных операций в программе на другой эквивалентный набор инструкций. Таким образом, производится внедрение информации без внесения искажений в алгоритм работы программы.

Современные методы стеганографии получают широкое применение в сфере защиты авторских прав. В объект авторского права может быть внедрена специальная метка, которая идентифицирует автора или законного получателя. Например, в каждую продаваемую копию программы будет внедрена метка, идентифицирующая лицензионного покупателя. В случае обнаружения пиратской копии программы, при помощи встроенной метки, без труда может быть отслежен пользователь, нарушивший лицензионное соглашение.

В данной статье предлагается улучшение метода стегоанализа текстовых данных, предложенного в работе [1]. Повышение эффективности стегоанализа достигается за счёт использования теста со смещением и применения теста хи-квадрат для различения двух различных распределений.

Структура статьи выглядит следующим образом. Во втором разделе даётся краткое описание используемого метода внедрения и существующих методов его обнаружения. В третьем разделе описывается предлагаемый метод стегоанализа и проводится экспериментальная оценка его эффективности. В четвёртом разделе происходит сравнение предлагаемого метода с известными аналогами.

2. Описание исследуемого метода стеганографии текстовых файлов

В предыдущей работе [1] мы подробно рассматривали различные виды методов внедрения, такие как синтаксические, генерирующие текст, подобный естественному, и семантические методы. В этой статье мы остановимся на семантическом методе внедрения Tyrannosaurus Lex (T-Lex), опубликованном в работе [2], использующем замену слов в предложении на их синонимы. Пример работы данного метода изображен на рис. 1.

So here we have an (0) unfixable / (1) unchangeable situation

Рис. 1. Пример работы программы T-Lex

В зависимости от выбранного синонима кодируется передаваемое сообщение. Предложение *So here we have an unchangeable situation* содержит стегосообщение − “1”. Данный метод требует наличия большого словаря синонимов.

Существует также обратная стеганографии задача – стегоанализ. Его целью является выявление факта наличия встроенного секретного сообщения в контейнере. В качестве критерия оценки эффективности методов стегоанализа используют вероятность обнаружения секретного сообщения в контейнере или вероятность возникновения ошибки. Существует два рода ошибок:

Ошибка I рода – случай, когда метод принимает пустой контейнер (без секретного сообщения) за заполненный (с секретным сообщением);
Ошибка II рода – случай, когда заполненный контейнер принимается за пустой.

Рассматривая вышеописанный метод внедрения можно отметить следующий недостаток, например, возможное нарушение стиля написания текста:

(0). . . *and make it still better, and say* ***nothing*** *of the bad – belongs to you alone.*
(1). . . *and make it still better, and say* ***nada*** *of the bad – belongs to you alone.*

Можно утверждать, что слово *nada* является нетипичным для использования некоторыми авторами, (в частности, *Jane Austen)*, что может вызывать подозрение. Указанный недостаток можно использовать в стегоанализе.

Опубликованный в [3] подход предполагает отслеживать нарушение семантических правил английского языка. Рассмотрим следующий пример: при встраивании сообщения в предложение *You use unlicensed software?* слово *unlicensed* может быть заменено на *unaccredited,* что не является нетипичным для английского языка. При определении текста, полученного программой *T-Lex,* ошибка I рода составляет 61.4 %. Ошибка II рода – 15.1 %. Стоит отметить, что данный уровень ошибок получается при анализе одного предложения. Следовательно, анализ текста, состоящего из нескольких предложений, будет более эффективным. Данный метод требует достаточно много времени.

Еще один метод предложен в статье [4]. Авторы предлагают использовать контекст слов для выявления факта внедрения. Например, под контекстом слова *intersect* размером четыре понимается два слова справа и слева (контекст выделен жирным шрифтом: *Synonym* ***sets do*** *intersect* ***with each*** *other*). Для указанного слова предлагается оценивать, насколько подходящим оно является для текущего контекста. Чтобы пояснить принцип оценивания, введём некоторые обозначения. Пусть w – оцениваемое слово (в нашем случае *intersect*). Тогда $f(w)$ – частота встречаемости слова в большом наборе текстов. $f(w, C)$ – частота слова w с контекстом C. Таким образом из двух слов a и b для заданного контекста C следует считать подходящим слово a, если выполняется одно из следующих условий:

1. $\big(f(a) = f(b)\big) \wedge \big(f(a, C) > f(b, C)\big)$, то есть при одинаковой частоте встречаемости слов a и b считают подходящим слово a, если оно встречается чаще с указанным контекстом C, чем слово b.

2. $\big(f(a, C) = f(b, C)\big) \wedge \big(f(a) > f(b)\big)$, при одинаковой частоте встречаемости слов a и b с учётом контекста C считают подходящим слово a, если оно чаще встречается в текстах (без учёта контекста).

3. $\big(f(a) < f(b)\big) \wedge \big(f(a, C) > f(b, C)\big)$, даже если слово a встречается реже слова b, но в заданном контексте C чаще используется слово a, то его считают подходящим. Например, в типичном тексте слово *unaccredited* может встречаться чаще, чем *unlicensed,* но при определенном контексте, как в случае с предложением *“You use unlicensed software”,* употребляется только *unlicensed*. Следовательно, в данном примере слово *unlicensed* является подходящим.

В табл. 1 показана эффективность работы метода. Мы видим, что указанный подход обеспечивает высокую точность работы при малом объёме входных данных.

Таблица 1. Вероятности ошибок работы метода, учитывающего контекст слова

Объём внедрения	Ошибка I рода	Ошибка II рода
20 бит	13.9 %	22.3 %
64 бит	7.8 %	8.9 %

Самым точным на сегодняшний день следует считать метод, опубликованный в работе [5]. Авторы предлагают оценивать, насколько случайным был выбор синонима в предложении. Очевидно, что в предложении со скрытой информацией выбор синонима выглядит более случайным. На больших наборах текстов (с внедрённым сообщением и без) производится сбор статистической информации, характеризующей выбор синонима, с использованием программы SVM [6]. Далее, анализируя любой подозрительный контейнер, по аналогичной статистической информации можно определить, к какому из текстов (с внедрённым сообщением или без) относится контейнер. Указанный метод работает с высокой эффективностью при малых объёмах входных данных. Ниже указан график зависимости точности работы метода от числа заменённых слов.

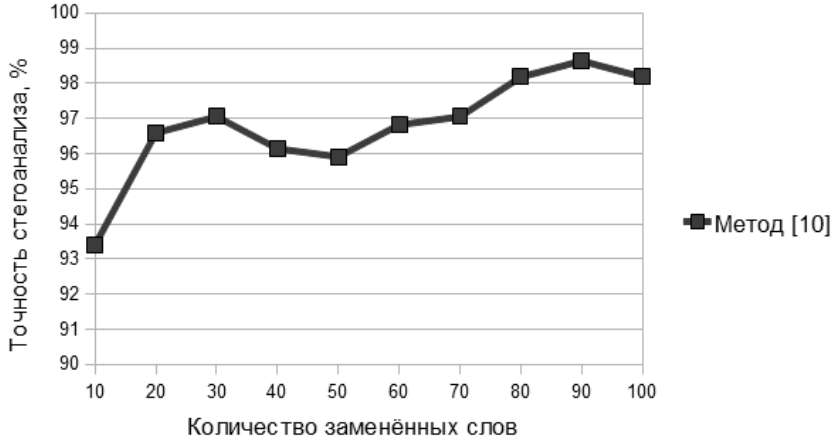

Рис. 2. Эффективность работы стегоанализа [10]

Для программы T-Lex принято считать, что замена каждого слова синонимом соответствует примерно одному биту внедрённого сообщения. Под точностью понимается отношение числа удачных попыток определения (пустых и заполненных контейнеров) к общему числу попыток. Мы видим, что даже при 10 заменённых словах метод работает достаточно эффективно.

3. Описание предлагаемого метода и результаты эксперимента

В данной работе предлагается модифицировать схему стегоанализа, предложенную в работе [1]. Напомним, что в предыдущем методе анализа предлагалось извлекать сообщение и разбивать его на элементы по L бит. Затем проверялась случайность распределения этих элементов. Таким образом, проверяя на случайность извлечённое сообщение, можно определить наличие факта внедрения. В работе [1] было установлено, что при $L = 4$ стегоанализ проходит наиболее эффективно.

В целях повышения эффективности работы метода предлагается модифицировать тест хи-квадрат путём подбора квантиля, дающего оптимальные результаты стегоанализа. Можно сказать, что квантиль представляет собой границу, отделяющую значения хи-квадрат случайных сообщений от неслучайных сообщений, и эту границу мы будет подбирать эмпирически. На рис. 3 показан график значений хи-квадрат, полученных от 100 сообщений, извлечённых из пустого контейнера, и от 100 сообщений, извлечённых из заполненных. Для наглядности значения были отсортированы по возрастанию и убыванию соответственно.

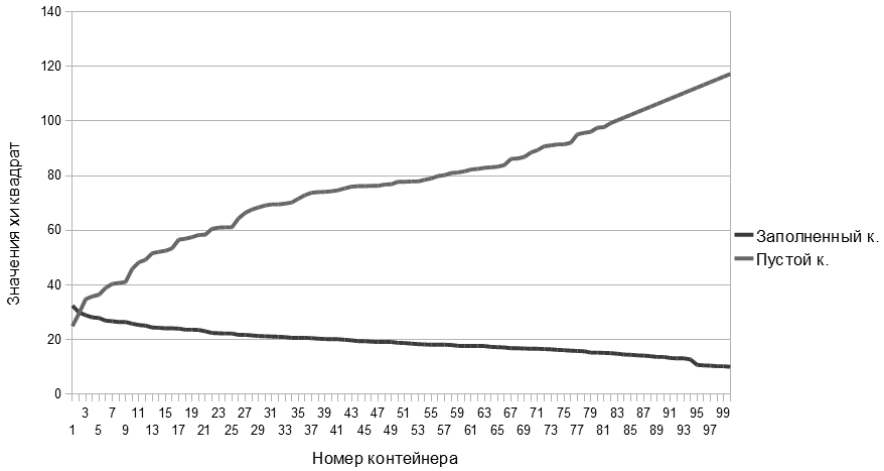

Рис. 3. Значения хи-квадрат, полученные для пустых и заполненных контейнеров

Мы видим, что кривые пересекаются. Это означает, что часть сообщений, извлечённых из пустых контейнеров, статистически неотличимо от сообщений из заполненных контейнеров. Следовательно, предлагаемый метод стегоанализа будет ошибаться в ряде случаев. Подбирая значения квантилей, можно уменьшать ошибку I рода, при этом будет увеличиваться ошибка II рода, и наоборот. Далее в эксперименте мы будем подбирать значение квантиля эмпирическим путём для повышения эффективности работы метода.

В настоящей статье для проверки случайности входного сообщения также предлагается использовать «тест со смещением», представленный на рис. 4.

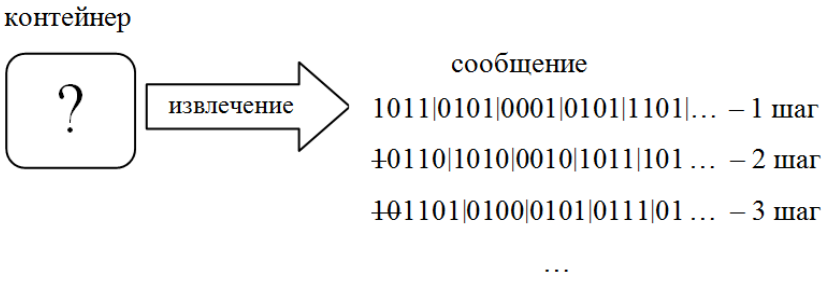

Рис. 4. Схема работы теста со смещением

Тест состоит из пяти шагов. На первом шаге извлечённое сообщение (начиная с первого бита) разбивается на элементы размером 4 бита и проверяется случайность распределения этих элементов. На каждом последующем шаге происходит отбрасывание одного левого бита сообщения. Затем производится аналогичное разбиение на элементы и проверка на случайность. Если на одном из пяти шагов сообщение признается неслучайным, то считается, что контейнер был пустым.

Теперь перейдём к описанию эксперимента, изображённого на рис. 5. На первом этапе определяется точность работы метода на пустых контейнерах. Набор текстов общим размером 400 Мб, состоящий из художественных произведений [7] на английском языке, объединялся в один файл (*Text*). Затем, из этого файла при помощи программы T-Lex извлекалось сообщение. Полученную двоичную последовательность (*message*) разбивали на отдельные фрагменты (*frag*). Количество фрагментов в нашем эксперименте составляет не менее 1000 шт. Далее, как показано на рис. 5, каждый такой фрагмент анализировался с помощью нашей программы стегоанализа. На заключительном шаге подсчитывались ошибки работы метода (ошибки I рода).

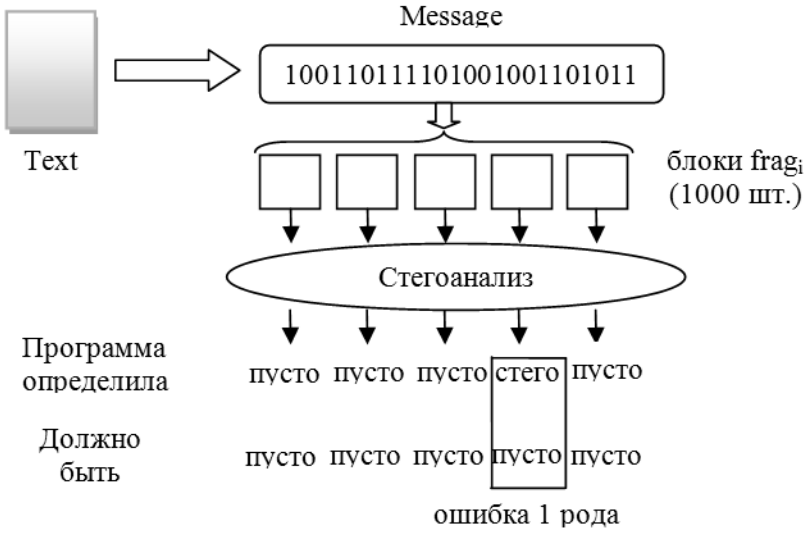

Рис. 5. Схема проведения эксперимента

На втором этапе эксперимента определялась точность работы метода на заполненных контейнерах. Передаваемое секретное сообщение генерировалось при помощи криптографического алгоритма RC6. Аналогично с первым этапом эксперимента полученное сообщение (*message*) разбивается на отдельные фрагменты (*frag*). Затем каждый фрагмент независимо от других анализируется с помощью нашей программы стегоанализа. На заключительном шаге подсчитываются ошибки работы метода (ошибки II рода). По результатам проведения эксперимента была построена табл. 2.

Таблица 2. Результаты проведения эксперимента при различных размерах фрагмента и значениях квантиля

Квантиль	Ошибка	Размер фрагмента (*frag*), бит									
		4000	3600	3200	2800	2400	2000	1600	1200	800	400
28	I рода	1.0	1.5	2.5	4.9	7.9	11.7	17.8	28.0	48.9	74.7
	II рода	5.1	4.9	5.8	6.3	5.7	4.6	4.3	5.5	6.1	4.5
33	I рода	2.3	3.0	5.2	9.2	12.3	18.8	27.6	45.8	67.7	83.8
	II рода	1.9	1.7	1.1	1.7	1.7	1.2	1.2	1.1	1.9	1.5
45	I рода	9.4	12.7	16.2	21.2	29.3	43.5	66.5	78.5	86.1	92.8
	II рода	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0

Подытожим результаты. В случае когда используемый квантиль равен 33, ошибки (I и II рода) стеготеста составляют 2.3% и 1.9% соответственно. При квантиле, равном 45, ошибка II рода практически отсутствует.

4. Сравнение метода с известными аналогами

Целью этой работы было создание стеготеста, обнаруживающего наличие внедрения в текстовые контейнеры. В табл. 3 проведено сравнение нашего метода с известными аналогами.

Таблица 3. Результаты сравнения метода с известными аналогами

Метод	[10]	[9]	[8]	Новый
Ошибка I рода	2.3%	7.8%	61.4%	2.3%
Ошибка II рода	0.5%	8.9%	15.1%	1.9%
Объём вх. данных	100-150 бит	64 бит	1-4 бит	4000 бит

Мы построили стеготест, имеющий достаточно простую реализацию, что положительно влияет на скорость анализа и высокую эффективность работы. Однако, как показано в табл. 3, к недостатку метода следует отнести достаточно большой объём входных данных.

Литература

1. Нечта И. В. Метод стегоанализа текстовых данных, основанный на использовании статистического анализа // Вестник СибГУТИ. 2011. №3. С. 27–34.
2. Winstein K. Tyrannosaurus Lex 1999. // URL: `http://alumni.imsa.edu/~keithw/tlex/` (дата обращения: 12.12.2009).
3. Taskiran C., Topkara U., Topkara M., Delp E. Attacks on Lexical Natural Language Steganography Systems // Proceedings of the SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VI. San Jose. 2006.
4. Yu Z., Huang L., Chan Z., Li L., Zhao X., Zhu Y. Steganalysis of Synonym-Substitution Based Natural Language Watermarking // International Journal of Multimedia and Ubiquitous Engineering. Vol. 4, No. 2, April, 2009.
5. Chen Z., Huang L., Yang W. Detection of substitution-based linguistic steganography by relative frequency analysis // Digital investigation 2011. Vol. 8(1), Elsevier. pp. 68–77.
6. Chang C., Lin C. Libsvm A library for support vector machines // URL: `http://www.csie.ntu.edu.tw/~cjlin/libsvm/` (дата обращения: 12.12.2009).
7. Gutenberg Project // URL: `http://www.gutenberg.org/wiki/Main_Page` (дата обращения: 13.06.2011).

Статья поступила в редакцию 28.11.2011;
переработанный вариант 22.03.2012

Нечта Иван Васильевич
аспирант, ассистент кафедры прикладной математики и кибернетики СибГУТИ,
тел. (383)2-698-272, e-mail: `www@inbox.ru`

Applying Statistical Methods for Secret Message Detection in Text Data

I. V. Nechta

In this paper we propose a method of improving the previous steganalysis. The essence of the original method was that extracted from the text secret message was tested randomness by using a chi-square test. The randomness of the extracted message meant that secret message was embedded. In this paper we propose a modified chi-square test so that the quartile used in the test was chosen empirically. This modification reduces the probability of steganalysis mistakes.
Keyword: Tyrannosaurus Lex, steganography, steganalysis, stegotext.