УДК 003.26

Метод стегоанализа текстовых данных, основанный на использовании статистического анализа [1]

И. В. Нечта

В данной статье рассматривается метод стегоанализа текстовых данных. Предлагается новый подход, позволяющий с достаточно высокой долей вероятности определить наличие встроенной скрытой информации в тексте. Суть метода состоит в том, что извлекаемое из текста сообщение проверяется на случайность при помощи теста хи-квадрат. Случайность извлечённого сообщения означает наличие внедрения.

Ключевые слова: стеганография, стегоанализ, стеготекст.

1. Введение

Стеганография – это наука о передаче «секретного» сообщения таким образом, чтобы сам факт передачи секретного сообщения оставался в тайне от стороннего наблюдателя. Слово стеганография дословно переводится как «тайнопись». В отличие от криптографии, ограничивающей доступ к передаваемому сообщению с помощью ключа, задача стеганографии состоит в сокрытии самого факта передачи секретного сообщения. Одним из развивающихся направлений является цифровая стеганография, позволяющая встраивать скрытое секретное сообщение в цифровые данные, такие как файлы звука, видео, изображения, текста и программного обеспечения. Сам факт передачи файлов, скажем, по сети Интернет не будет вызывать никакого подозрения. Цифровая стеганография получила широкое применение в сфере защиты авторских прав. Например, в случае с программным обеспечением внедрённое секретное сообщение является своеобразной «меткой», по которой может быть идентифицирована как продаваемая копия программы, так и её автор. Очевидно, что такая «метка» должна быть незаметна для постороннего и устойчива к удалению. Файл, в который внедряется секретное сообщение, будем называть *контейнером*.

На рис. 1 рассматривается классическая задача стеганографии. Алиса и Боб – участники обмена сообщениями. Их задача состоит в том, чтобы создать потаённый канал связи для передачи секретных сообщений. Ева может перехватывать подозрительные сообщения с целью выявления факта внедрения (это задача стегоанализа). Как показано на рис. 1, с помощью специальных алгоритмов Алиса встраивает секретное сообщение в контейнер и передаёт его Бобу. Благодаря этим алгоритмам стеганографии, Ева, перехватив сообщение, не сможет однозначно утверждать о наличии внедрения в контейнер. Таким образом, задача сокрытия факта передачи секретного сообщения будет выполнена. Более подробно методы стеганографии рассматриваются в монографии [1].

[1] Работа выполнена в рамках НИР по Гос. контракту № 02.740.11.0396 и грантом РФФИ №09-07-00005

Рис. 1. Классическая задача стеганографии

В статье предлагается новый метод стегоанализа текстовых данных, базирующийся на использовании статистического анализа.

2. Обзор существующих методов стеганографии текстовых файлов

Существующие методы встраивания сообщений в текстовые данные можно условно разделить на три вида:

Синтаксические методы. К таким методам можно отнести, например, предложенный в работе [2], использующий дополнительные пробелы между словами. Один пробел соответствует нулю, два – единице. Данный метод может широко применяться в файлах формата HTML (интернет страниц), поскольку наличие пробелов никак не влияет на отображение страницы. Недостатком можно считать лёгкую обнаруживаемость, так как обычно, при написании текста, дополнительные пробелы не используются. Существует возможность использовать специальные символы вместо пробелов, не отображающиеся в часто используемых текстовых редакторах. Ещё один метод, предложенный в работе [2] использует синтаксические ошибки при написании слов, например:

“This is the end”

“This iz the end”

Во втором варианте допущена опечатка. Наличие опечатки в определённых словах (в частности “iz”) означает, что бит передаваемой информации равен нулю, а отсутствие − единице. Таким образом происходит передача информации в тексте. Данный метод не является легко обнаруживаемым, так как в обычно тексте ошибки также могут встречаться.

Методы, генерирующие текст, подобный естественному. Рассмотрим метод, предложенный в работе [2], использующий контекстно-свободные грамматики для генерации естественно подобного текста. Правила:

S→ A B C
A→ She (0) | He (1)
B→ likes (0) | hates (1)
C→ milk (0) | apples (1)

В зависимости от значения бита передаваемого сообщения выбираем правило раскрытия нетерминального символа. Соответственно, если необходимо закодировать сообщение “101” получится: “He likes apples”. На сегодняшний день наиболее популярными стегосистемами являются Nicetext [3], Texto [4] и Markov-Chain-Based [5], так как имеют высокое соотноше-

ние размера входного сообщения к размеру генерируемого текста и получающийся текст максимально похож на естественный. Стоит отметить, что стеготекст, как правило, является бессмысленным.

Семантические методы. К этой группе относят **Tyrannosaurus Lex (T-lex)**, опубликованный в работе [6], использующий замену слов в предложении на их синонимы. Пример работы программы приведен на рис. 2.

	(0) decent		(0) metropolis
Tobolsk is a		little	
	(1) fine		(1) town

Рис. 2. Пример работы программы T-lex

В зависимости от выбранного синонима кодируется передаваемое сообщение. Предложение “Tobolsk is a decent little town” содержит стегосообщение − “01”. Данный метод требует наличия большого словаря синонимов. К недостатку таких методов относят возможное нарушение стиля написания текста. Например:

(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.*

Слово “*nada*” является не типичным для использования некоторыми авторами, в частности, Jane Austen, что может вызывать подозрение. Также существует метод, опубликованный в работе [7], преобразующий обычный текст в стеготекст путем перефразирования предложений. Например:

(0) *The caller identified the bomber as Yussef Attala, 20, from the Balata refugee camp near Nablus.*

(1) *The caller named the bomber as 20-year old Yussef Attala from the Balata refugee camp near Nablus.*

Данный метод обладает высокой степенью скрытности. Ещё один метод предложен в работе [8] и строится на том, что способов перевода предложения с одного языка на другой может быть несколько. Как показано на рис. 3, в зависимости от выбранного перевода предложения внедряется сообщение.

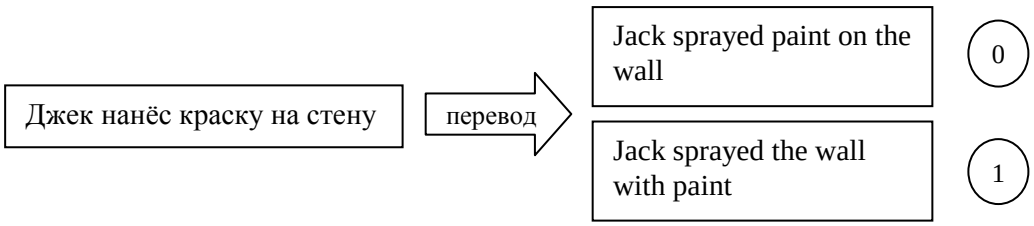

Рис. 3. Схема работы метода TBS

Выявить наличие стеготекста в этом случае также будет затруднительно. Фактически здесь имеет место обычное перефразирование.

3. Обзор существующих методов стегоанализа текстовых файлов

Как уже упоминалось ранее, существует обратная стеганографии задача – стегоанализ, целью которого является выявление факта наличия встроенного секретного сообщения в контейнер. В качестве критерия оценки эффективности методов стегоанализа используют вероятность обнаружения секретного сообщения в контейнере или вероятность возникновения ошибки. Существует два рода ошибок:

ошибка 1-го рода – случай, когда метод принимает пустой контейнер (без секретного сообщения) за заполненный (с секретным сообщением);

ошибка 2-го рода – случай, когда заполненный контейнер принимается за пустой.

Как уже было отмечено, использование методов, генерирующих текст, подобный естественному, имеет один недостаток − получается бессмысленный текст. Задача определения осмысленности текста требует участия человека. Однако, учитывая большой объём передаваемых сообщений в сети, это не всегда возможно. Поэтому особенно актуальна задача создания эффективных средств компьютерного анализа, работающая без участия человека.

На сегодняшний день существует большое число различных методов компьютерного стегоанализа. Рассмотрим их более подробно. Синтаксические методы можно обнаружить с помощью обычного анализа. Наличие большого количества орфографических ошибок в тексте будет вызывать подозрение. С другой стороны, если подобный стеготекст будет находиться не в файле, а встречаться в сетевых чатах или электронных форумах, то в таком случае наличие ошибок не будет являться подозрительным.

Методы, генерирующие текст, являются более трудно обнаруживаемыми. Как уже было отмечено, получаемый стеготекст всегда будет удовлетворять правилам грамматики языка. Для такого вида методов существуют следующие подходы. Один из них, предложенный в работе [9], использует частоту встречаемости слов и её дисперсию в анализируемом тексте. По полученным данным с помощью SVM классификатора определяется факт наличия стеготекста, сгенерированного программными средствами [3], [4] или [5] в контейнерах размером 5 Кб и более. Сумма ошибок 1-го и 2-го рода не превосходят 7.05 %.

При малых размерах входных данных наиболее эффективным является метод, опубликованный в работе [10], строящий модель языка текстов. После чего по имеющимся моделям стеготекста, обычного текста и текста контейнера, используя SVM классификатор, определяется, является ли подозрительный текст обычным или искусственным (стеготекстом). Точность обнаружения текста, сгенерированного программой [3], составляет 99.61 % на текстовых сегментах размером 400 байт и более.

В статье [11] предлагается новый метод, основанный на подходе, предложенном в работе Рябко Б.Я. [12], отличающийся от других тем, что для выявления факта наличия «стеготекста» используется сжатие обычным архиватором. Идея подхода состоит в том, что внедряемое сообщение нарушает статистическую структуру контейнера, повышая его энтропию. Следовательно, заполненный контейнер будет «сжиматься» хуже, чем незаполненный. В отличие от предыдущих методов, данный метод обладает рядом преимуществ. Анализ занимает сравнительно мало времени (порядка 0.1−0.5 сек на современных персональных компьютерах). Для проведения анализа не требуется словарей синонимов или правил грамматики языка, занимающих большой объём памяти. При обнаружении стеготекста, полученного программой [4], ошибки 1-го и 2-го рода составляют 0.02 % и 0.01 % соответственно.

Теперь рассмотрим стегоанализ наиболее трудно обнаруживаемых методов – семантических. Трудность обусловлена тем, что стеготекст не только не содержит ошибок, но также является осмысленным. Соответственно, даже анализ (определение бессмысленности) с участием человека не выявит наличие стеготекста. Однако такой вид внедрения также не лишен недостатков. В работе [13] опубликован стегоанализ, использующий недостатки семантических методов. При замене слов на их синонимы существует вероятность нарушения правил семантики языка. Например, при встраивании сообщения в предложение «What

time is it?» слово *time* может быть заменено на *period* или *duration*, что не является корректным для английского языка. При определении текста, полученного программой **[6],** ошибка 1-го рода составляет 61.4 %. Ошибка 2-го 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 работы.

Стегоанализ методов, базирующихся на переводе предложений, предлагается в статье [14]. Идея состоит в следующем. Во многих языках и, в частности, в английском языке существует группа слов, которые встречаются чаще других. Было обнаружено, что после внедрения стегосообщения в текст количество таких слов в тексте становится меньше. Пример изображён на рис. 4. Эффективность работы метода приведена в табл. 1.

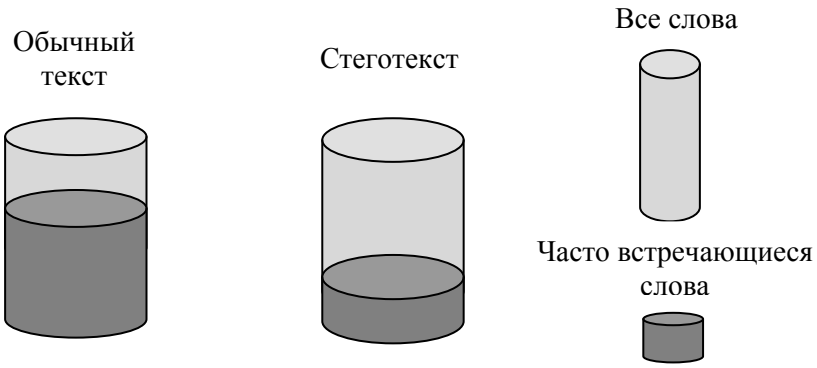

Рис. 4. Схема стегоанализа методов, базирующихся на переводе предложений

Как отмечают авторы, к недостаткам можно отнести необходимость знать систему стегоперевода (программу, с помощью которой внедрялось сообщения).

Таблица 1. Эффективность работы метода обнаружения внедрения, основанного на переводе предложений

Размер контейнера	Ошибки при обнаружении	
	1-го рода	2-го рода
20 Кб	10.7 %	13.9 %
40 Кб	5.4 %	7.8 %

В данной работе предлагается новый метод стегоанализа текстовых данных, полученных с помощью программы [6]. В отличие от предыдущих подходов, анализирующих содержимое контейнера, новый метод извлекает сообщение из контейнера и проверяет его на случайность. Как показывают экспериментальные данные, подход обладает высокой степенью эффективности.

4. Описание метода и результаты

Перед внедрением передаваемое стегосообщение будет зашифровано. Известно, что зашифрованное сообщение выглядит как случайная последовательность. Следовательно, извлекаемое из контейнера стегосообщение тоже будет выглядеть как случайное. В ходе экспериментов было установлено, что сообщение, извлечённое из пустого контейнера, выглядит менее случайным, чем зашифрованное сообщение. Таким образом, извлекая и анализируя сообщение из контейнера, мы сможем определить факт наличия внедрения. В нашем методе

случайность определяется с помощью теста хи-квадрат. На рис. 5 представлена схема работы предлагаемого метода стегоанализа. Из контейнера извлекается сообщение. Далее его разбивают на элементы размером L бит. Для анализа используется N таких элементов.

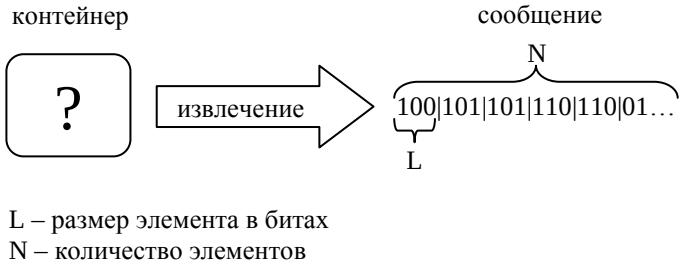

Рис. 5. Схема предлагаемого метода стегоанализа

В отличие от предыдущих известных подходов, предлагаемая схема является относительно простой и не требует каких-либо словарей или баз данных хранящих статистические характеристики текстов, необходимые для анализа.

Теперь определим эффективность предлагаемого метода экспериментально. Для этого была сформирована выборка контейнеров, состоящая из текстовых файлов (художественные произведения на английском языке [15]) общим размером в 150 Мб. Затем извлекалось стегосообщение до заполнения контейнера и после. Далее применялся тест хи-квадрат для выявления случайности распределения элементов (длинны L). При заполнении контейнеров (необходимых для проведения нашего эксперимента) рассматривалось два случая:

- внедрённое сообщение − это естественный текст на английском языке;
- внедрённое сообщение − это зашифрованное сообщение (мы будем имитировать его случайной последовательностью).

Рассмотрим случай, когда внедрено сообщение, являющееся естественным текстом на английском языке. Было установлено, что в этом случае сообщение, извлечённое из пустого контейнера, выглядит более случайным, чем извлечённое из заполненного. В табл. 2 приведены результаты 400 попыток определения наличия стеготекста предложенным методом.

Таблица 2. Результаты работы стеготеста, если внедрённое сообщение является естественным текстом на английском языке

	Ошибка 1 и 2 рода при различных N, в %															
	N=1000		N=700		N=500		N=300		N=100		N=70		N=30		N=15	
Род ошибки	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
L=1	58	30	44	47	34	61	28	77	14	15	10	15	8	16	6	100
L=2	72	14	59	33	45	52	35	74	14	93	12	68	9	23	49	100
L=3	83	9	64	23	49	42	34	65	17	93	7	27	87	3	0	100
L=4	97	5	95	21	86	42	68	68	36	7	92	3	38	37	0	100
L=5	84	1	68	9	50	24	33	51	88	15	92	26	0	100	0	100

Из приведённых результатов видно, что метод работает лучше всего при N = 30 и L = 1.

Теперь рассмотрим случай, когда внедрённое сообщение является случайной последовательностью. Данная ситуация рассматривается потому, что обычно при передаче секретного сообщения его перед внедрением шифруют. Мы будем имитировать зашифрованное сообщение последовательностью, полученной из генератора случайных чисел. В таком случае

сообщение, извлечённое из пустого контейнера, выглядит менее случайным, чем извлечённое из заполненного. Результаты работы стеготеста представлены в табл. 3.

Таблица 3. Результаты работы стеготеста, если внедрённое сообщение является псевдослучайной последовательностью

	Ошибка 1 и 2 рода при различных N, в %															
	N=1000		N=700		N=500		N=300		N=100		N=70		N=30		N=15	
Род ошибки	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
L=1	42	6	56	7	66	6	72	6	86	6	90	4	92	5	94	19
L=2	28	5	41	4	55	7	65	4	86	5	88	5	91	63	51	65
L=3	17	9	36	4	51	5	66	4	83	6	93	6	13	97	100	0
L=4	3	5	5	7	14	6	32	7	64	29	8	94	62	57	100	0
L=5	16	3	32	1	50	7	67	5	12	92	8	88	100	0	100	0

Из представленных результатов видно, что стеготест работает лучше всего при N = 1000 и L = 4.

Подытожим результаты. Величина ошибки (1 и 2 рода) работы стеготеста составляет 8 % и 16 % соответственно, когда контейнер заполняется естественным текстом на английском языке. Если же контейнер заполняется зашифрованным сообщением, то ошибки (1 и 2 рода) работы стеготеста составляют 3 % и 5 % соответственно.

5. Заключение

Целью этой работы было создание стеготеста, обнаруживающего внедрение в текстовые контейнеры. В данной статье рассматривалось два типа контейнеров − заполненных естественным текстом и псевдослучайной последовательностью. В случае естественного текста контейнер признаётся заполненным, если извлечённая последовательность выглядит (признаётся тестом хи-квадрат) как неслучайная. Для случайной последовательности, наоборот, контейнер признаётся заполненным, если извлечённая последовательность выгладит как случайная. К недостаткам предложенного метода можно отнести большой объём входных данных. В отличие от предыдущего метода [13], анализирующего одно предложение, представленный стеготест требует больший объём входных данных.

Литература

1. Рябко Б. Я., Фионов А. Н. Криптографические методы защиты информации: учебное пособие для вузов. М.: Горячая линия–Телеком, 2005. 229 с.
2. James C. Steganography Past, Present, Future // URL: `http://www.sans.org/reaing_room/whitepapers/steganography/steganogaphy_past_present_future_552.pdf` (дата обращения: 12.12.2009).
3. Nicetext // URL: `ftp://ftp.eenet.ee/pub/FreeBSD/distfiles/nicetext-0.9.tar.gz` (дата обращения: 12.12.2009).
4. Texto // URL: `http://www.nic.funet.fi/pub/crypt/steganography/texto.tar.gz` (дата обращения: 12.12.2009).

5. Markov-chain // URL: http://www.eblong.com/zarf/markov/chan.c (дата обращения: 12.12.2009).
6. Winstein K. Tyrannosaurus lex 1999. // URL: http://alumni.imsa.edu/~keithw/tlex/ (дата обращения: 12.12.2009).
7. Barzilay R., Lee L. Learning to paraphrase: An Universal approach using multiple-sequence aligment // URL: http://www.aclweb.org/anthology/N/N03/N03-1003.pdf (дата обращения: 12.12.2009).
8. Grothoff C., Grothoff K., Alkhutova L., Stutsman R., Atallah M. Translation-based steganography // In Proceedings of Information Hiding Workshop (IH 2005). Springer-Verlag. 2005. P. 15.
9. Chen Z., Huang L., Yu Z., Zhao X., Zheng X. Effective Linguistic Steganography Detection // IEEE 8th International Conference on Computer and Information Technology Workshops. 2008. P. 224–229.
10. Meng P., Huang L., Chen Z, Yang W., Li D. Linguistic Steganography Detection Based on Perplexity // URL: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?tp=&arnumber=5089098&isnumber=5089035 (дата обращения: 12.12.2009).
11. Нечта И. В. Эффективный метод стегоанализа, базирующийся на сжатии данных // Вестник СибГУТИ, 2010. № 1. С. 50–55.
12. Ryabko B. Compression-based methods for nonparametric density estimation, on-line prediction, regression and classification for time series // 2008 IEEE Information Theory Workshop. Porto, Portugal. 2008.
13. 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.
14. Meng P. Analysis and Detection of Translation-Based Steganography // LNCS. Springer-Verlag. V. 6387/2010. P. 208–220.
15. Gutenberg Project // URL: http://www.gutenberg.org/wiki/Main_Page (дата обращения: 13.06.2011).

Статья поступила в редакцию 16.06.2011;
переработанный вариант 04.07.2011

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

A method of steganalysis of text data based on statistical analysis

I. Nechta

This article is about a steganalysis method of text data. It proposes a simple and effective approach to detect the presence of embedded information in text files. A message extracted from the text is checked for randomness by means of chi-square test. The method is based on the fact that message randomness reveals the presence of embedding.

Keywords: steganography, steganalysis, stegotext.