Иван В. Нечта
Сибирский государственный университет телекоммуникаций и информатики,
ул. Кирова 86, Новосибирск, 630102, Россия
e-mail: ivannechta@gmail.com, http://orcid.org/0000-0003-0361-2742

НОВЫЙ МЕТОД СТЕГОАНАЛИЗА ТЕКСТОВЫХ ДАННЫХ, ПОЛУЧЕННЫХ КОДИРОВАНИЕМ ДЛИН СЕРИЙ СИНОНИМОВ

DOI: http://dx.doi.org/10.26583/bit.2018.2.10

Аннотация. В статье предложен новый метод стегоанализа, выявляющий текст, полученный методом кодирования длин серий синонимов. Анализируемый метод внедрения позволяет сохранять некоторые статистические свойства текста без изменений после внедрения скрытого сообщения. В частности, неизменными остаются: распределение вероятностей бит извлекаемого сообщения и распределение вероятностей использования синонимов текста, что обеспечивает высокую степень скрытности рассматриваемого метода внедрения. В ходе исследования было показано, что внедряемое сообщение изменяет статистическую структуру контейнера, и этот факт используется при стегоанализе. Разработанный стеготест сравнивает распределение вероятностей серий бит (с длиной не более 5 бит) в извлечённом из контейнера сообщении с эталонными распределениями, соответствующими пустому и заполненному контейнеру. Эталонные распределения были получены путём анализа 1000 контейнеров естественного текста, взятых из библиотеки Gutenberg Project. В работе рассматриваются два подхода к получению эталонных распределений. Первый подход предполагает анализ статистики сообщения, извлечённого из контейнера обычным способом (с помощью программы Tyrannosaurus Lex). Второй подход предполагает дополнительное преобразование сообщения в соответствии с анализируемым алгоритмом кодирования длин серий. Экспериментальные результаты позволяют утверждать о большей эффективности первого подхода. В качестве меры близости двух вероятностных распределений используется мера Кульбака-Лейблера. Показано, что реализованный метод позволяет обнаруживать наличие внедрения в контейнере с числом синонимов равным 500, при этом ошибка 1 рода равна 1.5%, ошибка 2 рода – 1.3%. По сравнению с известными аналогами предлагаемый метод имеет более высокую точность анализа при меньшем объёме входных данных.

Ключевые слова: стегоанализ, метод замены синонимов, tyrannosaurus lex.
Для цитирования. НЕЧТА, Иван В.. НОВЫЙ МЕТОД СТЕГОАНАЛИЗА ТЕКСТОВЫХ ДАННЫХ, ПОЛУЧЕННЫХ КОДИРОВАНИЕМ ДЛИН СЕРИЙ СИНОНИМОВ. Безопасность информационных технологий, [S.l.], n. 2, p. 114-120, 2018. ISSN 2074-7136. Доступно на: <https://bit.mephi.ru/index.php/bit/article/view/1118>. Дата доступа: 06 тау 2018. doi:http://dx.doi.org/10.26583/bit.2018.2.10.

Ivan V. Nechta
Siberian state university of telecommunications and informatic sciences,
Kirova st., 86, Novosibirsk, 630102, Russia
e-mail: ivannechta@gmail.com,, http://orcid.org/0000-0003-0361-2742

New method of steganalysis for text data obtained by synonym run-length encoding

Abstract. In this article, we present a new stegoanalysis method for detecting a text obtained by the synonym Run-Length Encoding. The analyzed RLE-method allows us to keep some statistical properties of the text after a secret message embedding. In particular, the probabilities distribution of the bits in the extracted message and the probabilities distribution of using text synonyms keep unchanged, that ensures a high secrecy degree of the considered embedding method. In this paper we show that the embedded message changes the probabilities distribution

of bit-series lengths in the extracted message, and this fact is used for our stegoanalysis. It was shown that the embedded message breaks the statistical structure of the container, and this fact is used for the stegoanalysis. The constructed stegotest compares the probability distribution of runs (with length no more than 5 bits) in the message extracted from the container with reference distributions corresponding to an empty and embedded containers. Reference distributions were obtained by analysing of 1000 natural-text containers taken from the Gutenberg Project library. In this paper we consider two approaches for obtaining reference distributions. The first approach deals with analyzing the statistic of the message extracted from the container in the usual way (using the Tyrannosaurus Lex program). The second approach involves an additional decoding of the message in accordance with the analyzed run-length encoding algorithm. Experimental results allow us to assert that the first approach is more effective. The Kullback-Leibler measure is used as a divergence measure of two probability distributions. It was shown that the proposed method makes it possible to detect presence of the secret message in the container with a number of synonyms equal to 500, while false negative error is 1.5% and false positive error is 1.3%. In comparison with the known analogs, the proposed method demonstrates higher accuracy of analysis for a smaller size of input data.

Keywords: steganalysis, synonym substitution method, tyrannosaurus lex.

For citation. NECHTA, Ivan V.. New method of steganalysis for text data obtained by synonym run-length encoding. IT Security (Russia), [S.l.], n. 2, p. 114-120, 2018. ISSN 2074-7136. Available at: <https://bit.mephi.ru/index.php/bit/article/view/1118>. Date accessed: 06 may 2018. doi:http://dx.doi.org/10.26583/bit.2018.2.10.

Введение

Классическая проблема стеганографии заключается в организации скрытого канала связи для обмена секретными сообщениями. Задача передачи скрытых данных впервые описана в работах Симмонса [1] и состоит в следующем. Пусть имеются два участника обмена сообщениями: Алиса и Боб. Их задача состоит в организации скрытой передачи данных под видом обычного обмена сообщениями. Постороннее лицо: Ева, анализирующая передаваемые сообщения, не должна заподозрить существование такого скрытого канала передачи данных. Алиса при помощи стеганографических алгоритмов встраивает секретное сообщение в безобидный на внешний вид объект данных, так называемый *контейнер*. Сам факт передачи контейнера по открытому каналу связи не является для Евы чем-то подозрительным. Боб, получив контейнер, сможет извлечь и прочитать секретное сообщение. Свойства стеганографических алгоритмов таковы, что Ева, подвергнув контейнер анализу, не сможет однозначно утверждать ни о наличии, ни об отсутствии факта внедрения скрытого сообщения.

На сегодняшний день известны различные методы стеганографии, использующие в качестве контейнера изображение, видео, аудио файлы и текст. Настоящее исследование посвящено методам текстовой стеганографии. Рассмотрим более подробно существующие методы внедрения скрытых данных в текстовые контейнеры, которые можно условно разделить на два класса.

Синтаксические методы. К данному классу принадлежат методы, например, описанные в работе [2], встраивающие в текст неотображаемые символы, дополнительные пробелы, которые в последствии не влияют на отображение текста. Существуют методы, например [3], использующие опечатки в определенных местах предложения. Данный класс методов является легко обнаружимым и в настоящее время активно не используются.

Семантические методы. В данный класс входят методы перефразирования предложений [4-5], в которых меняется форма их записи (активный, пассивный залог). Существуют методы, базирующиеся на переводе текста с одного языка на другой, например [6-7], в которых выбирается один из правильных вариантов перевода соответствующий скрываемому сообщению.

Известны методы генерации естественноподобных текстов по правилам контекстно-свободных грамматик языка, представленные в работе [8], такой текст является бессмысленным, но трудно обнаружимым (по сравнению с другими методами внедрения). Еще одним представителем данного класса является метод замены синонимов [9], в котором используется замена слов предложения на соответствующий скрываемому биту синоним. Смысл текста не меняется, но существует возможность нарушения устойчивых идиоматических выражений, что является недостатком метода.

Противоположной к задаче стеганографии является стегоанализ, заключающийся в выявлении факта передачи секретного сообщения. Считается, что синтаксические методы наиболее уязвимы к стегоанализу, так как легко обнаруживаются отклонения от грамматических правил языка в тексте. Семантические методы, напротив, являются более устойчивыми. Тем не менее, известно множество эффективных статистических методов анализа. Например, искусственно сгенерированный текст программой Texto [10] успешно обнаруживается в результате статистического анализа с помощью обычного архиватора, что описано в работе [11]. Известны методы, базирующиеся на статистических моделях n-грамматик (пар, троек словосочетаний), позволяющие проводить эффективный стегоанализ любых генерируемых текстов.

Метод замены синонимов успешно выявляется с помощью стегоанализа SVM-классификатором, представленного в работе [12]. Указанный метод базируется на анализе статистики встречаемости слов (синонимов) с определенным контекстом предложения. Для уменьшения статистических различий текста до и после внедрения, авторами статьи [13] была предложена модификация алгоритма замены синонимов. Во-первых, синоним выбирается с учётом вероятности его встречаемости в текстах английского языка. Во-вторых, встраивание бита скрытых данных осуществляется не в каждый синоним, а 0.3 бита на один синоним (в среднем), что понижает количество изменений исходного текста и, соответственно, нарушений его статистической структуры. В-третьих, известно, что распределение вероятностей бит сообщения, взятого из пустого и заполненного контейнера, при внедрении предыдущим методом [9] отличаются (что используется в стегоанализе [14]). Указанный алгоритм [13] позволяет сохранить соотношение изменённых нулевых и единичных бит. Согласно проведенным экспериментам стеготексты, полученные данным алгоритмом, практически не обнаруживаются при стегоанализе (ошибка при обнаружении стегоконтейнера достигает 90.27%).

В данном исследований предлагается метод стегоанализа текстов, полученных кодированием длин серий синонимов [13]. Анализируемый алгоритма внедрения использует длины серии бит (последовательность равных бит) для встраивания скрытых данных. Если серия имеет четную длину – внедрён ноль, нечетную – единица. Таким образом, меняя длину серий, встраивается секретное сообщение. Однако алгоритм имеет одну особенность, которая заключается в том, что при внедрении две стоящие подряд серии единичной длины удаляются (сливаясь в единую серию). Следовательно, стоит ожидать, что распределение вероятностей серий единичной длины у пустого и заполненного контейнера могут различаться. В настоящем исследовании реализуется данная идея для стегоанализа.

Описание предлагаемого метода. В рамках данного исследования рассматриваются два подхода к стеготесту. Оба варианта работают по одному и тому же алгоритму, представленному ниже, и отличаются только способом извлечения сообщения. Рассмотрим более подробно алгоритм стегоанализа.

Алгоритм 1. Алгоритм стегоанализа

Шаг 1. Извлечение сообщения из контейнера.

Шаг 2. Разбиение сообщения на элементы и расчет статистики.

Шаг 3. Классификация.

На первом шаге для извлечения сообщения из анализируемого контейнера мы используем исходную программу Tyrannosaurus Lex [9]. После извлечения сообщения получим *промежуточную* последовательность, состоящую из символов алфавита $A = \{0,1\}$. В первом подходе, мы используем для анализа *промежуточную* последовательность. Во втором подходе мы декодируем *промежуточную* последовательность в соответствии с алгоритмом кодирования длин серий [13] (назовем полученную последовательность *конечной*) и анализируем её.

Вторым шагом мы разбиваем извлечённое из контейнера сообщение на элементы. Предварительно введем следующее определение. *Серией* называется битовая последовательность $R = b_1, b_2, \dots, b_N : b_1 = b_2 = \dots = b_N;\ b \in \{0,1\}$. Здесь длина серии равна N.

В качестве элементов используются серии длины N, где $N \in [1; 5]$. После разбиения сообщения на элементы рассчитаем распределение их вероятностей. При обнаружении серии с длиной $N > 5$, она отбрасывалась и не учитывалась.

Для выполнения Шага 3 мы будем сравнивать полученное распределение со специальными эталонными распределениями, соответствующими пустому и заполненному контейнеру. Если полученное распределение «ближе» к эталонному распределению пустого контейнера, то анализируемый контейнер также признается пустым, в противном случае – признается заполненным. В качестве меры близости двух вероятностных распределений будем использовать меру Кульбака-Лейблера [15]. В ходе эксперимента использовались другие меры, но они оказались менее эффективны, чем упомянутая. Поэтому здесь будет рассмотрена только одна мера.

Далее рассмотрим процесс получения эталонных распределений. В качестве пустых контейнеров были подготовлены текстовые файлы, полученные из текстов архива Gutenberg Project [16], в количестве 1000 штук. Затем из контейнера извлекалась и анализировалась битовая последовательность в соответствии с Шагом 1 и 2 предлагаемого Алгоритма 1. Размер анализируемой последовательности составлял 1000 бит. Из полученных распределений вероятностей (обозначим их $Q_j = \{q_{j,1}, \dots, q_{j,N}\}$) вычислялось среднее арифметическое распределение (обозначим его $P = \{p_1, \dots, p_N\}$) согласно формуле 1, которое затем нормировалось. Нормировка необходима для выполнения требования о том, чтобы сумма вероятностей распределения равнялась единице. Таким образом, было получено эталонное распределение вероятностей, соответствующего пустому контейнеру. Здесь i - соответствует индексу вероятности элемента в распределении, j – соответствует номеру файла.

$$p_i = \frac{\sum_{j=1}^{1000} q_{j,i}}{1000} \quad (1)$$

Для получения эталонного распределения вероятностей, соответствующего заполненному контейнеру, производилось внедрение секретного сообщения в контейнер. Так как передаваемое сообщение предварительно шифруется и известно, что зашифрованное сообщение должно выглядеть неотличимым от истинно случайной последовательности (т.е. вероятности его бит равны 0.5 и между их появлением отсутствуют какие-либо закономерности), то мы имитировали секретное сообщение последовательностью, полученной из генератора псевдослучайных чисел.

Внедрение скрытого сообщения производилось по предложенному в работе [13] алгоритму. Далее сообщение обрабатывалось согласно Шагу 2 и 3 вышеописанного Алгоритма 1. В результате были получены следующие распределения вероятностей, представленные в таблице 1.

Таблица 1. Эталонные распределения вероятностей

Состояние	Анализируемая	Распределение вероятностей

контейнера	последовательность	
Пустой	промежуточная	{52.71, 24.39, 12.36, 6.77, 3.77}
Заполненный	промежуточная	{33.76, 35.58, 16.14, 9.63, 4.89}
Пустой	конечная	{55.16, 22.54, 12.05, 6.09, 4.16}
Заполненный	конечная	{51.62, 25.81, 12.90, 6.45, 3.22}

Из данных, представленных в таблице 1, видно, что количество серий в контейнере убывает с возрастанием её длины. Поэтому в тесте анализ проводится только по сериям с длиной не более 5 бит. Остальные серии при анализе не учитывались.

Анализ полученных результатов также позволяют утверждать, что стеготест на базе статистики промежуточной последовательности будет давать меньше ошибок, т.к. распределения вероятностей находятся «дальше» друг от друга. Для промежуточной последовательности (от пустого до заполненного) расстояние между распределениями вероятностей составляет 10.98, для конечной – 0.72. Далее мы будем рассматривать только один подход (на базе анализа промежуточной последовательности).

Экспериментальное исследование эффективности стегоанализа. Считается, что эффективность методов стегоанализа определяется его ошибками. Обычно используются следующие ошибки. Ошибка 1-го рода: случай, когда заполненный контейнер воспринимается как пустой. Ошибка 2-го рода: случай, когда пустой контейнер воспринимается как заполненный. Другой мерой ошибок, например, используемой в работе [13], является $Recall\ Rate\ (rr)$:

$$rr = \frac{tp}{tp+fn}, \tag{2}$$

где tp – количество правильно обнаруженных стегоконтейнеров, fn – количество стегоконтейнеров ошибочно распознанных как пустые контейнеры. Очевидно, что ошибка 1 рода (обозначим её как E_1) связана с $Recall\ Rate$ следующей зависимостью:

$$rr = 1 - E_1 \tag{3}$$

Для проведения экспериментальной оценки эффективности предложенного метода были отобраны тексты из архива Gutenberg Project [16], отличные от тех, которые использовались для получения эталонных распределений. Извлечение сообщения и заполнение контейнера проводилось с помощью программных средств, разработанных автором данной работы, по алгоритмам, описанным в предыдущей главе.

В результате были получены значения ошибок анализа 1000 контейнеров, представленные в таблице 2. В связи с тем, что объём внедрения в контейнеры одного размера может существенно отличаться (в зависимости от длины серий), то принято решение рассчитывать ошибку относительно длины *промежуточной* последовательности, что позволит осуществить объективную оценку результатов эксперимента.

Таблица 2. Результаты стегоанализа

Ошибки стегоанализа	Длина промежуточной последовательности, бит				
	100	200	300	400	500
Ошибка 1 рода	5.8%	3.9%	2.5%	1.4%	1.5%
Ошибка 2 рода	17.9%	6.1%	3.1%	1.6%	1.3%

Сравним полученные данные с результатами, представленными в работе [13]. Здесь используется SVM-классификатор, предложенный для стегоанализа текстовых данных в статье [12]. Согласно представленным авторами данным среднее количество синонимов в контейнере – 1738, что соответствует промежуточной последовательности размером в

1738 бит. Значение $Recall\ Rate$ на указанной длине составляет 9.73%, следовательно, ошибка 1 рода равна 90.27%.

Из анализа результатов, представленных в таблице 2 можно утверждать, что предлагаемый в данной работе стеготест существенно превосходит вышеупомянутый стегоанализ (SVM-классификатором) по эффективности, т.к. имеет меньше ошибок на более коротких анализируемых последовательностях.

Заключение.

В ходе работы был предложен метод стегоанализа текстовых данных, полученных кодированием длин серий синонимов [13]. Было установлено, что при внедрении указанным методом нарушается статистическая структура извлекаемого сообщения. В настоящей статье продемонстрировано, что для анализа статистических различий следует использовать промежуточную последовательность.

Полученные в результате эксперимента данные позволяют утверждать об эффективности предложенного метода стегоанализа. Реализованный алгоритм превосходит известные аналоги по точности обнаружения факта внедрения при меньших объёмах входных данных.

СПИСОК ЛИТЕРАТУРЫ:

1. Simmons G.J. The prisoners problem and the subliminal channel. In Advances in Cryptology Proceedings of Crypto 83. Plenum Press: 1984. P. 51-67.
2. Koluguri A., Gouse S., Reddy P. B. Text steganography methods and its tools. Int. J. Adv. Sci. Tech. Res. 2014. V. 2. No. 4. P. 888-902.
3. Judge J. C. Steganography: past, present, future. Lawrence Livermore National Lab., CA (US), 2001. – №. UCRL-ID-151879.
4. Atallah M. et al. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. Information Hiding. – Springer Berlin/Heidelberg, 2001. P. 185-200.
5. Meral H. M. et al. Natural language watermarking via morphosyntactic alterations. Computer Speech & Language. 2009. V. 23. No. 1. P. 107-125.
6. Grothoff C. et al. Translation-based steganography. International Workshop on Information Hiding. – Springer, Berlin, Heidelberg. 2005. P. 219-233.
7. Stutsman R. et al. Lost in just the translation. Proceedings of the 2006 ACM symposium on Applied computing. – ACM. 2006. P. 338-345.
8. Chapman M., Davida G. Hiding the hidden: A software system for concealing ciphertext as innocuous text. International Conference on Information and Communications Security. – Springer Berlin/Heidelberg, 1997. P. 335-345.
9. Winstein K. Lexical steganography through adaptive modulation of the word choice hash. URL: http://web.mit.edu/keithw/tlex/ (дата обращения: 20.01.2018).
10. Сайт программы «Texto». URL: http://www.nic.funet.fi/pub/crypt/steganography/texto.tar.gz (дата обращения: 20.01.2018).
11. Нечта И.В. Эффективный метод стегоанализа базирующийся на сжатии данных. Вестник СибГУТИ. 2010. №1. С. 50-55.
12. Xiang L. et al. Linguistic steganalysis using the features derived from synonym frequency. Multimedia tools and applications. 2014. V. 71. No. 3. P. 1893-1911.
13. Xiang L. et al. A novel linguistic steganography based on synonym run-length encoding. IEICE transactions on Information and Systems. 2017. V. 100. No. 2. P. 313-322.
14. Нечта И. В. Применение статистического анализа для обнаружения скрытых сообщений в текстовых данных. Вестник СибГУТИ. 2012. №. 1. С. 29-36.
15. Kullback S. Information Theory and statistics. — N. Y.: Dover Publications, 1997. — 399 p.
16. Сайт «Gutenberg Project». URL: http://www.gutenberg.org/wiki/Main_Page (дата обращения: 20.01.2018).

REFERENCES:

[1] Simmons G.J. The prisoners problem and the subliminal channel. In Advances in Cryptology Proceedings of Crypto 83. Plenum Press: 1984. P. 51-67.
[2] Koluguri A., Gouse S., Reddy P. B. Text steganography methods and its tools. Int. J. Adv. Sci. Tech. Res. 2014. V. 2. No. 4. P. 888-902.
[3] Judge J. C. Steganography: past, present, future. Lawrence Livermore National Lab., CA (US), 2001. – №. UCRL-ID-151879.
[4] Atallah M. et al. Natural language watermarking: Design, analysis, and a proof-of-concept implementation. Information Hiding. – Springer Berlin/Heidelberg, 2001. P. 185-200.
[5] Meral H. M. et al. Natural language watermarking via morphosyntactic alterations. Computer Speech & Language. 2009. V. 23. No. 1. P. 107-125.
[6] Grothoff C. et al. Translation-based steganography. International Workshop on Information Hiding. – Springer, Berlin, Heidelberg. 2005. P. 219-233.
[7] Stutsman R. et al. Lost in just the translation. Proceedings of the 2006 ACM symposium on Applied computing. – ACM. 2006. P. 338-345.
[8] Chapman M., Davida G. Hiding the hidden: A software system for concealing ciphertext as innocuous text. International Conference on Information and Communications Security. – Springer Berlin/Heidelberg, 1997. P. 335-345.
[9] Winstein K. Lexical steganography through adaptive modulation of the word choice hash. URL: http://web.mit.edu/keithw/tlex/ (дата обращения: 20.01.2018).
[10] Website «Texto». URL: http://www.nic.funet.fi/pub/crypt/steganography/texto.tar.gz (accessed 20.01.2018).
[11] Nechta I. V. Effective method of steganalysis based on the data compression. Vestnik SibSUTI. 2010. №1. P. 50-55. (in Russian).
[12] Xiang L. et al. Linguistic steganalysis using the features derived from synonym frequency. Multimedia tools and applications. 2014. V. 71. No. 3. P. 1893-1911.
[13] Xiang L. et al. A novel linguistic steganography based on synonym run-length encoding. IEICE transactions on Information and Systems. 2017. V. 100. No. 2. P. 313-322.
[14] Nechta I.V. Primenenie statisticheskogo analiza dlya obnaryjeniya skritih soobschenii v textovih dannih [Applying Statistical Methods for Secret Message Detection in Text Data]. Vestnik of SibSUTIS. 2012. №. 1. P. 29-36. (in Russian).
[15] Kullback S. Information Theory and statistics. — N. Y.: Dover Publications, 1997. — 399 p.
[16] Website «Gutenberg Project». URL: http://www.gutenberg.org/wiki/Main_Page (accessed 20.01.2018).

Поступила в редакцию – 23 марта 2018 г. Окончательный вариант – 04 мая 2018 г.
Received – March 23, 2018. The final version –May 04, 2018.