

ИСПОЛЬЗОВАНИЕ ОСОБЕННОСТЕЙ ФОРМАТА XML В МЕТОДАХ ТЕКСТОВОЙ СТЕГАНОГРАФИИ

**П.П. Урбанович, О.А. Нистюк, М.Г. Савельева,
Н.П. Шутько, А.Н. Николайчук**

*Белорусский государственный технологический университет,
ул. Свердлова, 13а, 220005 Минск, Беларусь,
для корреспонденции: shutko_bstu@mail.ru*

Описаны новые методы текстовой стеганографии, основанные на модификации пространственно-геометрических и цветовых параметров элементов электронных текстов-контейнеров и учете особенностей формата XML. Оценены пропускная способность и стойкость методов к модификации стеганоконтейнера.

Ключевые слова: XML-формат; текстовая стеганография; LSB-методы; модель RGB; стеганографическая стойкость.

USING OF THE XML FORMAT FEATURES IN THE METHODS OF TEXT STEGANOGRAPHY

**P.P. Urbanovich, O.A. Nistyuk, M.G. Saveleva,
N.P. Shutko, A.N. Nikolaichuk**

*Belarusian State Technological University,
Sverdlova str., 220005 Minsk, Belarus,
corresponding author: shutko_bstu@mail.ru*

New methods of text steganography, based on the modification of the spatial-geometric and color parameters of the elements of electronic text-containers and taking into account the features of the XML format are described. The covert channel capacity and resistance of the methods to the steganocontainer modification are characterized.

Keywords: XML format; text steganography; LSB methods; RGB model; steganographic resistance.

Введение

Как известно, передача и защита информации на основе стеганографии основана на сохранении в тайне самого факта реализации стеганографического преобразования. Это обстоятельство влияет на две взаимосвязанные цели исследований в данной предметной области: стеганографические методы должны обеспечивать высокую пропускную способность стеганоканала при максимально высоком уровне скрытности [1].

Указанные цели и соответствующие им направления разработки прикладных стеганометодов, как правило, основаны на такой модификации параметров стеганоконтейнера (при размещении тайной информации), которая сводила бы на нет эффективность визуальной атаки – с одной стороны, и не влияла бы на целостность осажденной (передаваемой) информации при случайной или преднамеренной модификации стеганоконтейнера. Если стеганоканал создается на основе электронных текстовых документов, то модифицировать можно как отдельные пространственно-цветовые параметры текста [2–5], так и отдельные атрибуты текстового файла-контейнера [6]. При этом уровень скрытности осаждаемого в контейнер сообщения, что нами отождествляется со скрытностью стеганоканала, связан с относительной частью модифицируемого параметра – по аналогии с относительной частью наименее значащих битов (*least significant bits*, *LSB*), используемых при размещении тайного сообщения, по отношению к битовой длине используемого параметра. Последний, например, при цветовой кодировке пикселей в одном цветом канале модели RGB составляет 8 битов.

В [7, 8] описаны общие концепции и основные особенности новых методов тестовой стеганографии, развивающих и дополняющих теорию и практику стеганографических преобразований текстовых документов-контейнеров на основе формата XML. В данной статье представлены новые результаты, характеризующие методы из [7, 8].

1. Основная часть

Использование XML-формата в текстовой стеганографии. Основополагающая идея использования пространственно-геометрических и цветовых параметров элементов текстовых документов в качестве носителей тайной информации базируется на специфике формата XML [5]. С его помощью решаются, в частности, задачи хранения и транспортировки данных в процессоре MS Word при обработке как векторной, так и растровой графики (текстовый документ также можно рассматривать как графический объект).

Файл формата *DOCX* представляет собой ZIP-архив, который содержит два типа файлов: файлы XML с расширениями *xml* и *rels* и медиафайлы (например, изображения). Можно сказать, что *DOCX*-файл представляет собой набор сжатых файлов формата XML, причем все текстовое содержимое электронного документа MS Word формата *DOCX* находится в одном файле – *document.xml*.

Для описания особенностей форматирования текста используются, как и в других языках разметки, теги. Например, тег описания свойств абзаца (`<w:pPr>`) содержит в себе вложенный тег описания межстрочного интервала, например, `<w:spacing w:lineRule="exact" w:line="360"/>`, который обозначает, что высота межстрочного интервала задана точно и составляет 18 пунктов (пт). Для описания форматирования отдельных символов используется тег `<w:rPr>`. Например, в конструкции `<w:rPr><w:sz w:val="28"/><w:szCs w:val="28"/></w:rPr>` параметр `<w:sz w:val="28"/>` измеряется в $\frac{1}{2}$ пт и в данном случае указывает, что кегль текста равен 14 пт.

Как следует из изложенного, размещение тайной информации в электронном текстовом документе может осуществляться путем модификации определенных параметров, отвечающих за форматирование текста. Далее рассмотрим важные особенности предлагаемых стеганометодов.

Метод на основе растровой графики и цветовой модели RGB. В качестве базового элемента контейнера, цветовые параметры которого модифицируются в модели *RGB* при размещении тайной информации, выступает пиксель изображения. Внедрение/извлечение информации происходит в пикселях, имеющих одинаковое значение (одно из 256) в одном или нескольких цветовых каналах (R, G, B).

Для внедрения сообщения *M* в контейнер *C* необходимо выбрать массив пикселей, для которых совпадает значение координат одного или двух цветовых каналов. Пояснение к этому дает рисунок 1, на котором в увеличении представлен фрагмент буквы. В изображениях с большим количеством полутонов, монохроматических или черно-белых изображениях выбор пикселей, в которых будет происходить внедрение, целесообразно осуществлять по двум цветовым каналам. При этом непосредственно для внедрения информации в выбранные пиксели используется один канал.

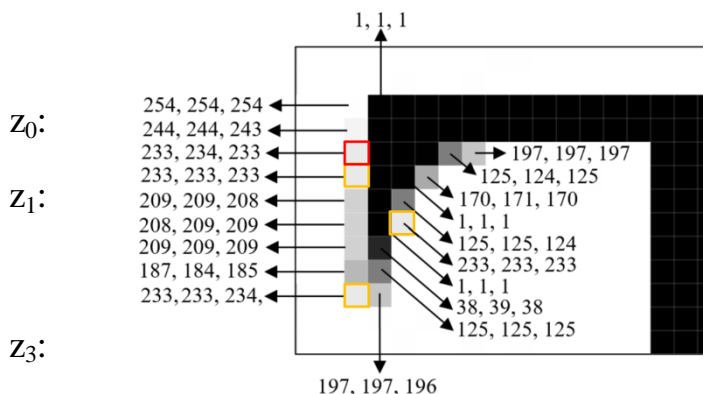


Рисунок 1 – Пояснение к алгоритму анализа и выбора пикселей массива *Z*

Важным шагом алгоритма внедрения является выбор массива пикселей, Z ($Z = \{z_i\}$, здесь $i = \overline{0, \text{length}(Z)}$). Необходимо также определить следующие элементы: c_{RGB} – цветовой канал с совпадающими цветовыми параметрами пикселя, $c_{RGB} \in R, G, B$, c_{RGB}' – цветовой канал для внедрения сообщения M , $c_{RGB}' \in R, G, B$, s_{jn} – пиксельный элемент документа C , $s_{jn} \in C$ ($C = \{s_{jn}\}$, $j = \overline{0, t}$, $n = \overline{0, r}$), t и r – размер C : соответственно ширина и высота в пикселях, φ – ключевое значение цветового кода канала c_{RGB} , $\varphi \in \{0, 1, \dots, 255\}$. Последний параметр используется для увеличения пропускной способности создаваемого скрытого канала. Для этого следует провести анализ того, в каком цветовом канале имеется больше пикселей с одинаковым значением цветового кода (c_{RGB}) и выбрать это значение в качестве параметра φ . Канал для внедрения (c_{RGB}') выбирается произвольно из оставшихся двух (или оба). Канал c_{RGB}' не должен использоваться при формировании массива пикселей Z . Например, после проанализированного фрагмента изображения-контейнера (часть буквы; рисунок 1) можно сказать, что $c_{RGB} = R$, $c_{RGB}' = G$ (как следующий после R), $\varphi = 233$.

Из массива Z выбирается базовый пиксель. Внедрение M будет происходить в канал c_{RGB}' при сравнении значений цветовых кодов канала c_{RGB}' пикселя для внедрения, $c_{RGB}'(z_0)$ ($z_i \in Z$) и базового пикселя, $c_{RGB}'(z_i)$. В этом примере базовый пиксель имеет цветовой код (233, 234, 233). Далее внедрение будет происходить при сравнении значений кода канала G базового (первого) и второго, базового и третьего и т. д.; в конечном итоге – базового и n -ного пикселей массива Z .

Для извлечения внедренного сообщения необходимо выбрать массив пикселей Z_D (пример показан на рисунке 2), где совпадает значение кода одного или нескольких цветовых каналов (по аналогии с формированием Z).

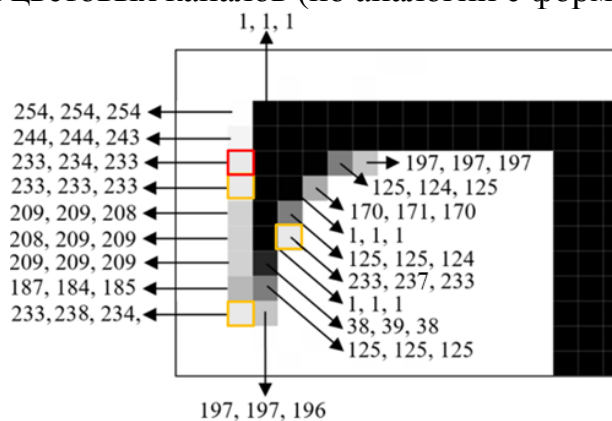


Рисунок 2 – Пример массива Z_D на фрагменте изображения

В отличие от массива Z в массив Z_D помещаются пиксели, для которых $c_{RGB}'(s_{jn})$ отличается от $c_{RGB}'(z_0)$ на $2Q$ единиц в любую сторону (диапа-

зон для выбора: $c_{RGB}'(z_0) - 2Q \leq c_{RGB}'(s_{jn}) \leq c_{RGB}'(z_0) + 2Q$). Параметр Q выбирается из условия: $4 < Q < 10$; чем шире граница указанного диапазона, тем выше пропускная способность канала, а указанные нижний и верхний пределы обеспечивают высокую устойчивость канала к визуальным атакам, что установлено многочисленными тестами. В качестве ключей стеганографического преобразования [4] может использоваться информация о том, какой канал (или несколько каналов) используется для выбора пикселей для внедрения M , координаты базового пикселя в массиве Z или алгоритм выбора базового пикселя из массива пикселей для внедрения, канал для внедрения, значение φ , значение Q .

Для оценки эффективности метода по параметрам пропускной способности и устойчивости к стеганоконтейнеру к модификациям разработано программное приложение, с помощью которого, в частности, установлено, что текстовый документ формата *PNG* с тайной информацией сохраняет целостность этой информации после конвертации в форматы *GIF*, *BMP*, *TIFF* (со сжатием и без сжатия), однако при конвертации в *JPG* примерно 70-75% битов исходного сообщения M являются ошибочными. Пропускная способность канала на основе метода примерно соответствует аналогичному параметру для семейства методов на основе LSB. Важным является то, что изображения с большим количеством полутонов, черно-белые и монохромные изображения будут обеспечивать сравнительно более высокую пропускную способность, так как они построены на основе большего количества пикселей с совпадающими кодами цветового канала.

Метод на основе модификации параметров контура символов текста является близким аналогом методов, основанных на модификации цвета символов текста, а также параметров апроса и кернинга [2, 3]. Параметры контура можно легко найти, выбрав пункт меню *Главная* в среде MS Word. К основным из таких параметров относятся: цвет, прозрачность, ширина, составной тип, тип штриха и др. Глубина изменения каждого из параметров влияет на пропускную способность и устойчивость преобразования к визуальным атакам в известном соотношении (лучше одно – хуже другое). Для примера на рисунке 3 показан вид букв с контуром и без контура. Даже при значительном увеличении контур остается визуально незаметным.



Рисунок 3 – Примеры шрифтового оформления с контуром и без него

При реализации метода необходимо принять во внимание символы, которые не могут быть дополнены контуром. К ним относятся: ", #, \$, %, &, ', (,), *, +, «,», -, ., /, :, ;, <, =, >, ?, @, [,], ^, _ , ` , {, |, }, ~, -, \|s, \.

Для анализа эффективности метода авторами создано отдельное приложение. Оригинальный текст с внедренным сообщением конвертировался в форматы *PDF*, *TXT*, *DOC* и обратно. После обратной конвертации сообщение *M* восстановить не удалось. Однако использование всех вышеперечисленных архиваторов не влияет на целостность *M* после распаковки архива.

Выводы

Для оценки эффективности (пропускной способности, устойчивости к случайным или преднамеренным модификациям стеганоконтейнера с размещенной тайной информацией) разработаны специальные программные средства, зарегистрированной в Государственном реестре информационных ресурсов РБ. Установлено, что метод на основе растровой графики и цветовой модели RGB обеспечивает целостность осажденной информации при конвертации стеганоконтейнера в большинство основных форматов, а метод на основе модификации параметров контура символов текста – при конвертации в форматы *PDF*, *TXT*, *DOC* и обратно. Пропускная способность канала на основе предложенных методов примерно соответствует аналогичному параметру для семейства методов на основе LSB.

Библиографические ссылки

1. Subramanian N., Elharrouss O., Al-Maadeed S., Bouridane A. Image Steganography: A Review of the Recent Advances // IEEE Access. 2021. № 9. P. 23409–23423. DOI: 10.1109/ACCESS.2021.3053998.
2. Shutko N., Urbanovich P., Zukowski P. A method of syntactic text steganography based on modification of the document-container aprosh // Przegląd Elektrotechniczny. 2018. № 94(6). P. 82-85. DOI:10.15199/48.2018.06.15.
3. Шутько Н.П. Алгоритмы реализации методов текстовой стеганографии на основе модификации пространственно-геометрических и цветовых параметров текста // Труды БГТУ. 2016. № 6. С. 160–165.
4. Urbanovich P., Shutko N. Theoretical Model of a Multi-Key Steganography System // Recent Developments in Mathematics and Informatics. Contemporary Mathematics and Computer Science. Vol. 2, Chapter 11. Lublin: KUL, 2016. P. 181–202.
5. Блинова Е.А., Сущеня А.А. Применение нескольких стеганографических методов для осаждения скрытых данных в электронных текстовых документах // Системный анализ и прикладная информатика. 2019. № 2. С. 32–38. URL: <https://doi.org/10.21122/2309-4923-2019-2-32-38>
6. Урбанович П.П., Юрашевич Д.Э. Использование системных свойств и парамет-

ров текстовых файлов в стеганографических приложениях // Теоретическая и прикладная криптография: материалы междунар. научной конференции. Минск, 20–21 октября 2020 г. Минск: БГУ, 2020. С. 68–73.

7. Нистюк О.А. Защита текстовой информации с помощью добавления контура к символам текста // Информационные технологии: материалы 86-й научно-техн. конф. профессорско-препод. состава, научных сотр. и аспирантов, Минск, 31.01 – 12.02 2022 г. Минск: БГТУ, 2022. С. 59–62.
8. Савельева М.Г. Метод стеганографического внедрения тайной информации в WEB-документы на основе растровой графики // Информационные технологии: материалы 86-й научно-техн. конф. профессорско-препод. состава, научных сотр. и аспирантов, Минск, 31.01 – 12.02 2022 г. Минск: БГТУ, 2022. С. 52–54.