V. S. Matveeva, A. V. Mamaev

Wavelet Analysis for Localization of Heterogeneities in Byte Distribution in a File to Searching for Encrypted Data

Key words: wavelet analysis, basis function, a criterion to searching for encrypted files

This article is focused on the choice of approach to localize a deviation in byte distribution in a file by means of the wavelet analysis. In the article the dependence of scale parameter and wavelet coefficients values is revealed. Experiments are conducted to confirm a hypothesis that encrypted data does not have deviations in the byte distribution. The results show that the false positive is reduced to zero and the proposed approach has its drawbacks, in the form of increased false negative; however it will be improved in subsequent works.

В. С. Матвеева, А. В. Мамаев

ВЕЙВЛЕТ-АНАЛИЗ ДЛЯ ЛОКАЛИЗАЦИИ НЕОДНОРОДНОСТЕЙ В РАСПРЕДЕЛЕНИИ БАЙТ В ФАЙЛЕ С ЦЕЛЬЮ ИДЕНТИФИКАЦИИ ЗАШИФРОВАННЫХ ДАННЫХ

В статье автора [1] показано, что для оценки статистических свойств содержимого файла f можно использовать преобразование: $S: f \rightarrow \{\rho_1, \rho_2, \ldots, \rho_L\}$, где

$$\rho_i = \frac{(\text{количество точек из файла в i}-\text{м фрагменте})}{(\text{количество всех точек в i}-\text{м фрагменте} = W * H)},$$

$i \in [1,L]$, L — количество получившихся фрагментов в результате прохождения по содержимому плоскости методом скользящего окна размером W x H.

В рамках тестирования форматов файлов с высокой энтропией, не говоря уже об остальных форматах файлов, было обнаружено, что в распределении плотностей имеются выраженные всплески, которые и предлагается отслеживать для выявления зашифрованных данных [1].

На основании такого отличия в распределении плотностей для файлов различных форматов, в том числе с высокой энтропией, предлагается проверить **гипотезу**: если файл не содержит выраженного отклонения от среднего значения плотности в распределении плотностей, то он может содержать данные, близкие к случайным.

Для определения отклонений будет применено вейвлет-преобразование.

Предлагаемый подход

Вид анализируемых распределений напоминает сигналы, для анализа которых, как правило, используются два подхода: преобразование Фурье (и его аналоги: преобразование Хартли, преобразование Уолша — Адамара и др.) и вейвлет-преобразование. Оба подхода являются разновидностями спектрального анализа, и их цель — разложение сигнала на составляющие частоты и амплитуды. Однако они имеют разные принципы работы.

Преобразование Фурье не позволяет получить по-настоящему локализованную информацию, что требуется в поставленной задаче, так как базисная функция является постоянно осциллирующей. Кроме этого, разложение сигнала производится только по частоте.

При этом вейвлет-преобразование осуществляет разложение сигналов во временном и в частотном пространстве одновременно. Элементом базиса вейвлет-преобразования являются функции, которые стремятся к нулю на $\pm\infty$ и чем быстрее, тем лучше [2]. В связи с этим базисные функции обладают свойством локализации.

Вейвлет-анализ состоит в расчете вейвлет-коэффициентов для функции, описывающей сигнал. Формулы расчета приведены в [2]. Вейвлет-коэффициенты W_{ab} показывают схожесть между анализируемым сигналом и анализирующим вейвлетом. Имеется в виду схожесть с точки зрения распределения частот.

В качестве анализируемого сигнала в рамках решаемой задачи выступает вектор плотностей $\{\rho_1, \rho_2, \ldots, \rho_L\}$ с дискретными значениями, поэтому формула для вейвлет-коэффициентов преобразуется следующим образом:

$$W_{ab} = \sum_{i=1}^{L} \rho_i * \psi_{ab}^{*}(i),$$

где L — длина проверяемой последовательности плотностей, ψ_{ab}^{*} — комплексное сопряженное базисной функции ψ_{ab}.

Движение по графику распределения плотностей осуществляется с помощью скользящего окна ширины a. Параметр сдвига b отвечает за позицию скользящего окна на гистограмме. Значение масштаба a принято выбирать степенью двойки, то есть 2, 4, 8, 16, 32, 64 и т. д. [2].

Как было сказано выше, базисная функция выбирается таким образом, чтобы ее форма походила на искомую форму отклонения, так как рассчитываемые коэффициенты показывают степень близости анализируемой последовательности и анализирующего вейвлета. На рис. 1–4 приведены самые известные базисные функции [3].

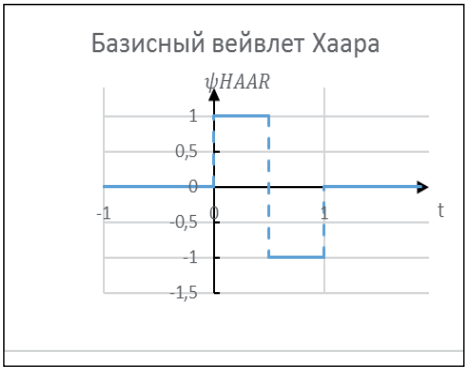

Рис. 1. График базисного вейвлета Хаара

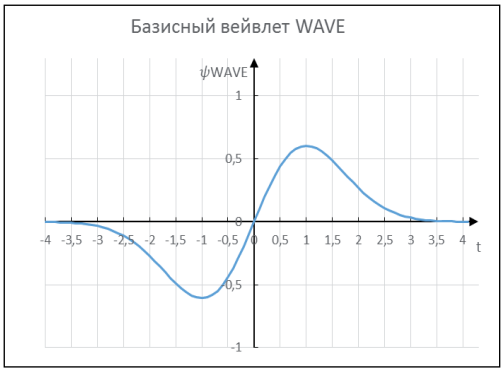

Рис. 2. График базисного вейвлета WAVE

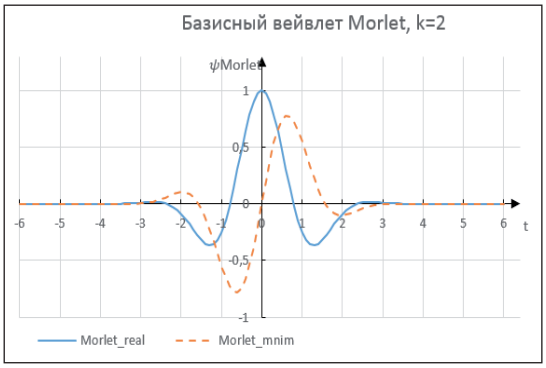

Рис. 3. График действительной и мнимой частей базисного вейвлета Морле при $k = 2$

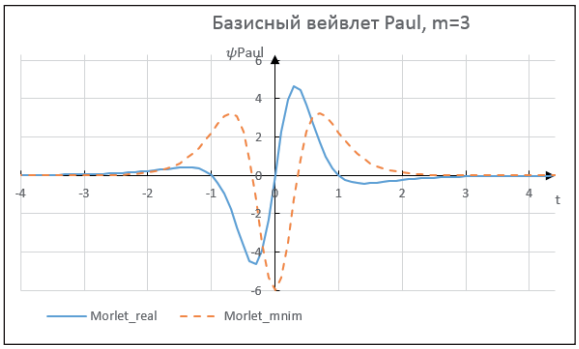

Рис. 4. График действительной и мнимой частей базисного вейвлета Пауля при $m = 3$

Все приведенные базисные функции способны локализовать определенный вид отклонения. Однако всплеск или спад в распределении плотностей необязательно будет представлять выраженное нарастание или последующий спад значений, которое может быть выявлено вейвлет-преобразованием на основе базисных функций на рис. 2–4. Отклонение может быть выражено в качестве единичного очень большого значения. Поэтому подходит поиск резкого перепада с помощью вейвлет-преобразования на основе базисной функции Хаара:

$$\psi_{HAAR}\ (\mathrm{t}) = \begin{cases} 1,\ 0 \le t < 1/2, \\ -1,\ 1/2 \le t < 1, \\ 0,\ t < 0,\ t \ge 1. \end{cases}$$

Так как эта базисная функция на концах принимает значение 0 и не имеет мнимой части, то вейвлет-преобразование вектора плотностей с использованием базисной функции Хаара может быть переписано следующим образом:

$$W_{ab} = \frac{1}{\sqrt{a}} \sum_{i=b}^{b+a} \rho_i * \psi_{HAAR}\left(\frac{i-b}{a}\right),$$

где a — параметр масштаба (то есть количество значений элементов последовательности ρ_{diff}, используемых для подсчета вейвлет-коэффициента), b — параметр сдвига (то есть позиция в последовательности ρ, от которой начинается отсчет элементов в этой последовательности), ρ_i — значение плотности i-го фрагмента. Проведя анализ приведенной формулы, можно сделать вывод, что при оценке вейвлет-коэффициентов имеет значение именно их амплитуда, а не знак.

Проверка гипотезы, приведенной во введении, производится в разделе «Эксперимент» путем подсчета диапазона значений вейвлет-коэффициентов для зашифрованных файлов и проверкой вейвлет-коэффициентов для файлов других форматов на выход их значений за пределы рассчитанного диапазона.

Зависимость значений вейвлет-коэффициентов от параметра масштаба

Несложно показать, что ширина окна a влияет на принимаемые значения вейвлет-коэффициентов и на детализацию анализа распределения. Масштабный коэффициент a обратно пропорционален частоте. Маленький масштаб соответствует большой частоте, а значит, более детальному рассмотрению сигнала, и наоборот. Большой масштаб расширяет сигнал, маленький — сжимает. Но в базисе масштабный коэффициент стоит в знаменателе, а значит, все рассуждения обратны.

На примере файла формата pdf (размер файла: 12 983 385, энтропия файла: 7,891518) произведем построение распределения плотностей для двух размеров окон: 100 x 100, 2000 x 5 и вейвлет-коэффициентов для разных значений параметра масштаба a, которые представляют собой степени 2: 2, 4, 8, 16, 32, 64, 128. На рис. 5–16 приведены некоторые из графиков.

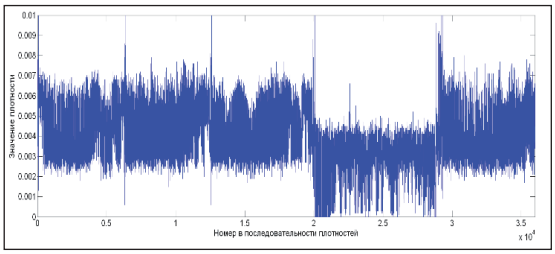

Рис. 5. Распределение плотностей для файла формата pdf и размера окна 100 x 100

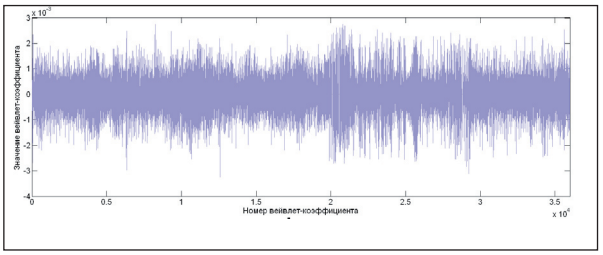

Рис. 6. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 100 x 100 и a = 2

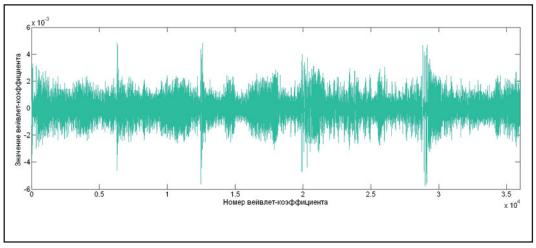

Рис. 7. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 100 x 100 и a = 8

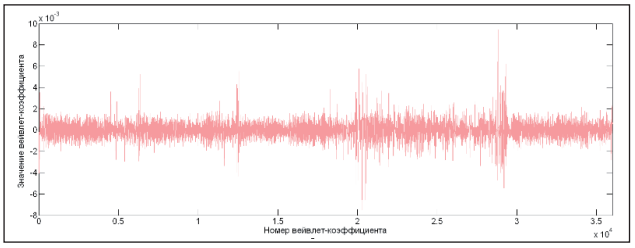

Рис. 8. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 100 x 100 и a = 32

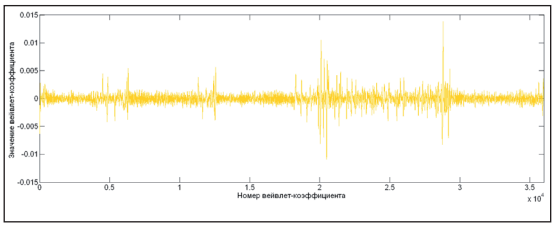

Рис. 9. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 100 x 100 и a = 64

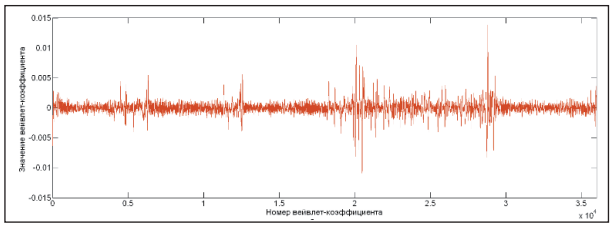

Рис. 10. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 100 x 100 и a = 128

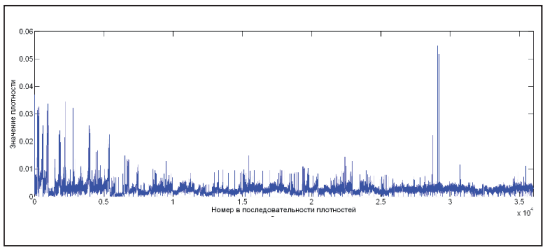

Рис. 11. Распределение плотностей для файла формата pdf и размера окна 2000 x 5

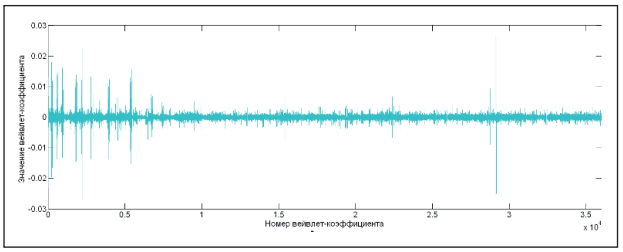

Рис. 12. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 2000 x 5 и a = 4

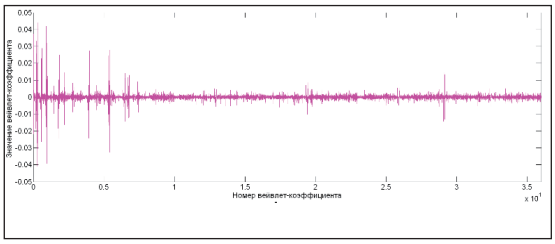

Рис. 13. Распределение вейвлет-коэффициентов файла формата pdf, размера окна 2000 x 5 и a = 16

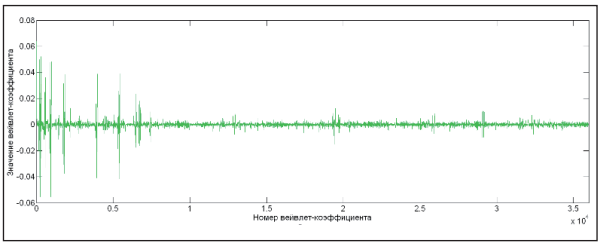

Рис. 14. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 2000 x 5 и a = 32

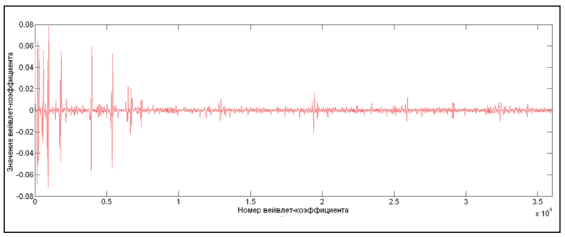

Рис. 15. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 2000 x 5 и a = 64

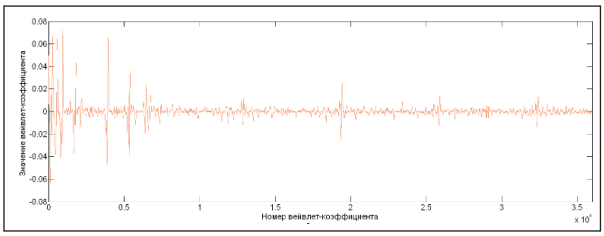

Рис. 16. Распределение вейвлет-коэффициентов для файла формата pdf, размера окна 2000 x 5 и a = 128

Из приведенных графиков можно сделать следующие выводы:

- локальные отклонения могут быть выявлены небольшим значением ширины окна, в то время как протяженные — большим;
- маленькие значения параметра $a = 2, 4, 8$ позволяют выявлять единичные отклонения, в то время как большие 16, 32, 64 — групповые.

Эксперимент

Сгенерируем коллекцию файлов с данными, зашифрованными по алгоритму AES, с помощью программного модуля «SharpAESCrypt» версии 1.0 на основании файлов из коллекций govdocs1 [4] и filetypes1 [5]. В полученной коллекции размеры файлов лежат в диапазоне: от 10 Кб до 911 Мб. Коллекция содержит 11 608 файлов.

Для нее произведем расчет диапазонов максимальных значений, принимаемых вейвлет-коэффициентами при выборе размеров окон: 100 x 100 и 2000 x 5, а также для значений параметра a = 2, 4, 8, 16, 32, 64, 128. Для каждой комбинации (размер окна, *a*) произведем расчет порогового значения, выше которого на тестовой коллекции зашифрованных файлов вейвлет-коэффициенты не принимают значение.

С учетом этих пороговых значений проделаем те же вычисления для нескольких коллекций:

1. Коллекция govdocs1 [4], в которой находятся часто встречающиеся форматы файлов: pdf, docx, csv, jpeg, jpg, pptx, rtf, swf, xlsx, gz и др. Коллекция содержит 795 221 файл.

2. Коллекция filetypes1 [5] файлов форматов: asp, avi, b64, b85, bz2, css, dll, elf, exe, ext3, fat, flv, jar, js, m4a, mov, mp3, mp4, ntfs, pst, rpm, rtf, random, swf, txt, tbird, url, wav, wma, xlsx и др. Коллекция содержит 1793 различных файла. В коллекции содержатся 445 файлов со случайными данными, которые в таблице 1 приведены отдельно.

3. Коллекция файлов, зашифрованных посредством программного обеспечения «BestCrypt» версии 8.25.3.1, «TrueCrypt» версии 7.1a. Таким образом, созданы файлы, зашифрованные по широко

применяемым алгоритмам: AES, Serpent, Twofish, Blowfish, 3DES, CAST, IDEA, RC6, ГОСТ 28147-89, имеющие размеры: 4 Кб, 512 Кб, 1 Мб, 512 Мб, 1 Гб. Минимальный выбранный размер объясняется ограничением используемых средств. Коллекция содержит 45 зашифрованных файлов.

4. Коллекция файлов-архивов формата RAR, защищенных паролем, которые созданы на основании коллекций govdocs1 и filetypes1 с помощью ПО «WinRAR» версии 5.01. Коллекция содержит 11 184 файла. Размеры файлов лежат в диапазоне от 20 Кб до 283 Мб.

В рамках эксперимента производился расчет максимальных значений вейвлет-коэффициентов для каждого файла и их сравнение с полученными пороговыми значениями. В случае превышения порогового значения в рамках тестирования файл признается незашифрованным.

Результаты тестирования приведены в таблицах 1, 2 и сравниваются с результатами тестирования критерия «Хи-квадрат на равномерность» (уровень значимости выбран 0,01) для тех же коллекций. В таблицах приводятся значения параметра масштаба *a*, при которых получены наименьшие значения ошибок.

Таблица 1. Ошибка первого рода, полученная в результате тестирования

Оцениваемая величина / Название коллекции	Количество файлов	Предлагаемый подход Ошибка I рода (100 x 100, a = 16)	Предлагаемый подход Ошибка I рода (2000 x 5, a = 128)	Хи-квадрат Ошибка I рода
filetypes1(random)	445	0 %	0 %	1,798 %
Зашифрованные файлы с помощью специального ПО «BestCrypt» и «True-Crypt» (без учета заголовков)	45	0 %	0 %	0 %
Зашифрованные файлы AES	11 608	0 %	0 %	1,081 %
Зашифрованные файлы-архивы формата RAR	11 184	0 %	0 %	1,829 %

Таблица 2. Ошибка второго рода, полученная в результате тестирования

Оцениваемая величина / Название коллекции	Количество файлов	Предлагаемый подход Ошибка II рода (100 x 100, a = 16)	Предлагаемый подход Ошибка II рода (2000 x 5, a = 128)	Пересечение результатов по двум размерам окон	Хи-квадрат Ошибка II рода
govdocs1	795 221	4,914 %	3,408 %	1,706 %	0,276 %
filetypes1	1295	8,853 %	7,336 %	6,926 %	5,778 %

Вывод

В работе предложен новый способ оценки распределения байт в файле на основании расчета плотности распределения точек на плоскости (номер байта в файле, значение байта). Проверяется гипотеза о том, что для зашифрованных файлов отклонение в распределении плотностей на этой плоскости не превосходит полученного экспериментально порогового значения. Проверка гипотезы основывается на выявлении отклонений с помощью вейвлет-анализа. В качестве базисной функции выбирается вейвлет Хаара. Кроме этого, в работе показывается зависимость между характером отклонения и параметром масштаба. На основании полученных зависимостей проведен эксперимент, в ходе которого предложенный подход выявления отклонений в распределении плотностей обнаруживает зашифрованные и случайные данные с нулевой ошибкой первого рода. В ошибку второго рода вносят вклад мультимедиафайлы и файлы-архивы размером до 1 Мб, а также файлы нераспространенного формата файлов-изображений jb2. Развитием подхода является уменьшение ошибки второго рода за счет подбора значения окна и параметра масштаба для файлов размером до 1 Мб.

СПИСОК ЛИТЕРАТУРЫ:

1. *Матвеева В. С.* Критерий оценки содержимого файлов различных форматов на предмет их близости к случайным данным // Безопасность информационных технологий. 2015. № 1 (в печати).
2. *Астафьева Н. М.* Вейвлет-анализ: основы теории и примеры применения // Успехи физических наук. 1996. Т. 166. № 11. С. 1145–1170.
3. *Farge M.* Wavelet Transforms and their applications to turbulence // Annual Review Fluid Mechanics. 1992. № 24. P. 395–457.
4. Govdocs1 – (nearly) 1 million freely-redistributable files [Электронный ресурс]: ресурс с тестовыми материалами для компьютерной криминалистики. Digital Corpora, 2014. URL: http://digitalcorpora.org/corpora/govdocs (дата обращения: 10.12.2014).
5. Index of /corp/files/filetypes1 [Электронный ресурс]: ресурс с тестовыми материалами для компьютерной криминалистики. Digital Corpora, 2014. URL: http://digitalcorpora.org/corp/files/filetypes1/ (дата обращения: 10.12.2014).

REFERENCES:

1. *Matveeva V. S.* Kriterij otsenki soderzhimogo fajlov razlichnyh formatov na predmet ih blizosti k sluchajnym dannym // Bezopasnost Informacionnyh Texnologii. 2015. № 1 (v pechati).
2. *Astafeva N. M.* Vejvlet-analiz: osnovy teorii i primery primeneniya // Uspexi fizicheskih nauk. 1996. T. 166. № 11. P. 1145–1170.
3. *Farge M.* Wavelet Transforms and their applications to turbulence // Annual Review Fluid Mechanics. 1992. № 24. P. 395–457.
4. Govdocs1 – (nearly) 1 million freely-redistributable files [Electronic resource]: a website of digital corpora for use in computer forensics education research. Digital Corpora, 2014. URL: http://digitalcorpora.org/corpora/govdocs/.
5. Index of /corp/files/filetypes1 [Electronic resource]: a website of digital corpora for use in computer forensics education research. Digital Corpora, 2014. URL: http://digitalcorpora.org/corp/files/filetypes1/.