РАСШИРЕНИЕ ГРАНИЦ ПРИМЕНЕНИЯ МЕТОДОВ ДЕШИФРОВАНИЯ ШИФРА ВИЖЕНЕРА

Бабаш А.В.[1], Гузовс Р.[2], Касаткин С.В.[3], Прохоров А.Н.[4], Слимов Н.А.[5]

Цель статьи: *ввести строгую формализованную модель содержания открытого текста, зашумленного текста, определить границу уровня зашумления открытого текста, при котором его содержание не может быть понято носителем языка.*

Метод: *дешифрование измененного шифра Виженера, в роли ключа в котором используется некачественная гамма – периодическая зашумленная последовательность.*

Полученный результат: *обоснована формализация задачи дешифрования шифра случайного гаммирования, шифра Виженера, приведено понятие содержания искаженного открытого текста. Разработаны алгоритмы выделения содержания открытого текста (ключевых слов), его замушления, восстановления содержания зашумленного текста с исправлением орфографических ошибок, с сегментацией. Введена шкала уровней понимания текста. На основе повторяющихся зашумлений текстов схожей тематики получено опытное значение среднетекстового уровня содержания восстановленных зашумленных текстов длин от 100 до 2000 символов.*

Ключевые слова: *шифрование Виженера, открытый текст, содержание открытого текста, шифрованный текст.*

DOI:10.21681/2311-3456-2019-5-42-50

1. Введение

Существует мнение о том, что шифр случайного гаммирования является недешифруемым шифром. Это мнение основано на известном и справедливом утверждении о его совершенстве (по К. Шеннону). Однако в работе [1] были высказаны сомнения: *«Теоретически существует совершенно секретный шифр (иными словами, абсолютно стойкий шифр), но единственным таким шифром является одна из форм так называемого одноразового шифрблокнота, в которой открытый текст комбинируется с полностью случайным ключом и имеющимся у нас алгоритмом такой же длины»*.. И далее, по всей видимости, авторы заметили подвох. Они выразили сомнение: «*ключи, выработанные с помощью некоторого датчика истинно случайных чисел, будут качественными с вероятностью, отличающейся от единицы на ничтожно малую величину*». Мы предполагаем, что под некачественным возможным ключом авторы имели в виду, в частности, сплошь нулевую ключевую последовательность (слабый ключ), при которой шифрованный текст совпадает с открытым текстом ОТ1 поданным на вход шифра. Такое сомнение мы считаем ошибочным, так как любой содержательный открытый текст ОТ на выходе шифра можно получить при значениях ключа ОТ-ОТ1. Действительно, ОТ+ОТ-ОТ1=ОТ- шифрованный текст. Тем не менее, данная ошибка авторов не снимает вопрос о нахождении метода дешифрования шифра случайного гаммирования.

Ниже решается задача дешифрования шифра гаммирования, использующего в качестве ключей почти периодические последовательности. В терминах шифра Виженера данные ключи могут трактоваться как искаженные ключи – лозунги, или как искаженная «локально периодическая последовательность» [1]. Данная задача равносильна задаче определения содержания зашумленного открытого текста по известному шифрованному тексту в шифре Виженера. Для решения этой задачи математически формализованы понятия: содержание открытого текста и его зашумление.

2. Общие сведения о шифрах случайного гаммирования и Виженера

2.1. Описание шифра случайного гаммирования

Пусть в алфавите используемого языка n *букв:* $\{i_0, i_1, i_2, \ldots, i_{n-1}\}$. Тогда обозначим $I = \{0,1,\ldots,n-1\}$ номера букв в данном алфавите. Обозначим через $X = I^L$ множество всех возможных текстов длины L используемого языка ($x \in X$), через $K = I^L$ множество ключей ($k \in K$), через $Y = I^L$ множество шиф-

1 Бабаш Александр Владимирович, доктор физико-математических наук, профессор НИУ ВШЭ, г. Москва, Россия. E-mail: ababash@hse.ru

2 Гузовс Рихард, студент магистратуры НИУ ВШЭ, г. Москва, Россия.

3 Касаткин Семен Владимирович, студент магистратуры НИУ ВШЭ, г. Москва, Россия.

4 Прохоров Арсений Николаевич, студент магистратуры НИУ ВШЭ, г. Москва, Россия.

5 Слимов Никита Алексеевич, студент магистратуры НИУ ВШЭ, г. Москва, Россия.

УДК 003.26.09 DOI: 10.21681/2311-3456-2019-5-42-50

рованных текстов ($y \in Y$). Также обозначим множество открытых содержательных текстов $M \subseteq X$. Для шифрования открытого содержательного текста буквы текста кодируются своими номерами. При расшифровании номера шифрованного текста декодируются в буквы. Для $x = x_1 x_2 \dots x_L \in M$, $k = k_1 k_2 \dots k_L \in K$, $y = y_1 y_2 \dots y_L \in Y$ уравнение шифрования имеет вид $\bar{x}_i + \bar{k}_i = \bar{y}_i \, mod \, n,$. Уравнение расшифрования: $n + y_i - k_i = x_i$. $modn$ Предполагается, что ключи выбираются случайно и равновероятно, и независимо от открытого текста.

2.2. Описание шифра Виженера (ШВ)

Шифр использует ключи, являющиеся повторением D -граммы $d = d_1 d_2 \dots d_D \left(d_i \in I\right)$ m раз и дополнением ключа до длины $L = m * D + r$:

$$k = k_1 k_2 \dots k_L \in K \quad k_{i+l*D} = d_i$$

$$i \in \{1, \dots, D\}, \; l \in \{0, \dots, \mathrm{m}+1\}, \; i + l * D \le L .$$

Шифрование проводится по правилам шифра случайного гаммирования. [1, 2]

3. Дешифрование шифра случайного гаммирования

Идея дешифрования ШСГ впервые была высказана в работе [1]. Она состояла в том, что определение открытого текста x по шифрованному текста y зашифрованного ШСГ ключом-гаммой k , близкой по метрике Хэмминга к периодической последовательности d сводится к дешифрованию искаженного открытого текста x гаммой d ШВ. То есть, дешифрование y с помощью методов дешифрования ШВ даст искаженный открытый текст x.

Таким образом, обоснована постановка и решение следующей задачи. Дан шифр Виженера и его ключ k . Ключ k искажается шумом. Искаженный ключ k' используется для шифрования сообщения x . Задача состоит в определении содержания открытого текста x по известному шифрованному $y = f_k(x)$.

Проблематика очистки от шума актуальна не только для содержательных текстов, но и для видеозаписей, изображений и аудиозаписей. В частности, в [12, 13] предлагаются решения для очистки речи, аудиозаписей. Задача удаления шума на изображениях требует анализа шаблонов возможных шумов. В [14-16, 20] рассмотрены проблемы выделения контуров объектов и удаления шума с изображений. Особенное внимание в [17-19, 21-25] уделено вопросам зашумлений в сигналах.

4. Формализация задачи дешифрования шифра случайного гаммирования (ШСГ)

Определение 1. Вектор n -шума для ключа $k = k_1 k_2 \dots k_L$ – любой вектор $\mu(j_1)\mu(j_2)\dots\mu(j_n)$, где $\mu(j_i) \in \{I\} \setminus \{0\}$, $i \in \{1, \dots, n\}, 1 \le \; < \; < \dots < \; \le$ $1 \le j_1 < j_2 < \dots < j_n \le L$.

Определение 2. Искаженным ключом k' ключа $k = k_1 k_2 \dots k_L$ вектором n -шума $\mu(j_1)\mu(j_2)\dots\mu(j_n)$ называется набор $k' = k_1' k_2' \dots k_L'$, где $k_s' = k_s \oplus \mu(s)$ для $\{j_1, \dots, j_n\}$ и $k_s' = k_s$ для $s \notin \{j_1, \dots, j_n\}$. .

Аналогично определению 2 вводится понятие искаженного открытого текста $x' = x_1 x_2' \dots x_L'$ вектором n -шума $\mu(j_1)\mu(j_2)\dots\mu(j_n)$ исходного открытого текста $x = x_1 x_2 \dots x_L$.

Из законов функционирования шифра Виженера вытекает следующее.

Утверждение 1. Задача дешифрования обобщенного шифра Виженера равносильна задаче дешифрования искаженных открытых текстов для шифра Виженера (т.е. без искажения ключа, но с искажением открытого текста).

Замечание. В [1] доказано следующее утверждение. Пусть $\left(p(i), i \in I\right)$ – вероятностное распределение букв алфавита I в открытых текстах. $x = x_1 x_2 \dots x_L$ выборка из указанного распределения. Тогда искаженный открытый текст $x' = x_1' x_2' \dots x_L'$ случайно и равновероятно выбранным вектором n -шума $\mu(j_1)\mu(j_2)\dots\mu(j_n)$ является выборкой из распределения $\left(p'(i), i \in I\right)$, где

$$p'(i) = \frac{n|I|}{L(L-n+1)(|I|-1)} + \frac{n}{L(L-n+1)(|I|-1)}$$

Данные выше утверждение и замечание позволяют нам перейти к решению задачи дешифрования искаженного открытого текста в шифре Виженера.

Решить эту задачу можно известными методами, в частности, методом, указанным выше. В связи с этим мы сосредоточим свое основное внимание на результатах такого дешифрования. А именно:

1) насколько сложно понять содержание искаженного открытого текста, полученного в результате дешифрования;

2) при какой мере зашумления открытого текста невозможно определить его содержание.

5. Формализация понятия содержания искаженного открытого текста

Содержание искаженного открытого текста определяется исходя из возможности восстановления содержания исходного открытого текста по искаженному. Для оценки будем использовать ответы на вопрос «много ли мы узнали из полученного искаженного текста». Предлагаются следующие уровни ответов:

- Почти все понятно: большая часть слов и слогов (от 66% до 100%) понятны, искаженный текст дает достаточно полное представление о смысле и содержании сообщения, незначительные детали сообщения непонятны;
- Средне понятно: около трети всех слов и слогов (от 33 до 65%) понятны, отдельные дешифрованные фрагменты дают представление о отдельных темах сообщения, содержание и характер всего сообщения непонятен;
- Плохо понятно: только некоторые слова и слоги понятны (от 0 до 33%), сам дешифрованный текст не связан, нелогичен и не осмыслен, понятны части слов и фраз, остальное разобрать невозможно.

Для получения оценок автоматически (без участия человека) была разработана авторская программа, использующая следующие подпрограммы:

- Подпрограмма, реализующая алгоритм $TF-IDF$ [7], для выделения ключевых слов текста;
- Авторская подпрограмма, зашумляющая исходный открытый текст $x = x_1 x_2 \dots x_L$ с помощью вектора n-шума $\mu(j_1)\mu(j_2)\dots\mu(j_n)$;
- Подпрограмма, исправляющая орфографические ошибки (спеллер) [10];
- Подпрограмма, реализующая разделение (сегментацию) слов, объединенных в одно слово в результате зашумления пробела между словами.

Рассмотрим ключевые особенности работы перечисленных подпрограмм.

5.1. Подпрограмма, реализующая алгоритм TF-IDF

Известный алгоритм $TF-IDF$ предназначен для определения значимости слова из открытого текста [7]. Значимость или веса слов, полученные в результате выполнения алгоритма, могут применены для классификации текстов, получения короткой выдержки из документов [8, 9]. При выполнении алгоритма в решаемой задаче для каждого слова w открытого текста $t \in T$ рассчитывается два показателя: $tf(w,t)$ и $idf(w,T)$.

Тексты t принадлежат коллекции T текстов схожей тематики.

$tf(w,t) = \dfrac{C(w,t)}{W(t)}$ – term frequency (частота слова), частота употребления слова w в рамках текста t, где $C(w,t)$ – число вхождений слова w в текст t, $W(t)$ – количество слов в тексте t.

$idf(w,T) = \log \dfrac{|T|}{\sum_{t_i \in T} C(w,t_i)}$ – inverse document frequency (обратная частота документа), инверсия частоты, с которой слово w встречается в текстах $t_i \in T$ коллекции T, $t \in T$.

Результат выполнения алгоритма $TF-IDF$ для одного слова w текста $t \in T$ есть величина $tfidf(w,t,T) = tf(w,t) * idf(w,T)$ [7]. Больший показатель $tfidf(w,t,T)$ получают слова с высокой частотой встречаемости в тексте $t \in T$ и низкой частотой встречаемости в других текстах $t_i \in T$.

5.2. Подпрограмма, зашумляющая исходный открытый текст

Авторская подпрограмма предназначена для зашумления открытых текстов $t \in T$ вектором n-шума $\mu(j_1)\mu(j_2)\dots\mu(j_n)$. Входными параметрами подпрограммы являются открытый текст t и величина относительного уровня зашумления $n_{\%} = \dfrac{n}{W(t)}$, где $W(t)$ – количество слов в тексте t. Результатом работы алгоритма является искаженный (зашумлённый) текст t'.

5.3. Подпрограмма, исправляющая орфографические ошибки

Используется подпрограмма Yandex.Speller, реализующая интерфейс HTTP API для исправления орфографических ошибок в тексте t', поданном на выход под программе. Выходными параметрами Yandex.Speller является список орфографических ошибок (с одним вариантом исправления для каждой ошибки) в словах текста t', после чего в текст t' вносят правки согласно полученным ошибкам и вариантам исправления. В результате работы подпрограммы получается исправленный текст t''.

В основе Yandex.Speller лежит библиотека машинного обучения CatBoost. С ее помощью спеллер может исправить такие слова как «адникассниеи» («одноклассники») [10]. Спеллер является программной реализацией неточного распознавания слов, естественного для носителя языка.

5.4. Подпрограмма, реализующая разделение (сегментацию) слов

Проблема, решаемая данной подпрограммой, – разделение соединенных в результате зашумления

пробелов между словами. Например, в результате зашумления два следующих друг за другом слова «король» и «говорит» превратились в «корольктоворит». Носитель русского языка с легкостью прочтет эти два слова и выделит смысл, однако для других подпрограмм такое сочетание букв будет отдельным словом. Результатом работы подпрограммы будет текст t''', над словами которой выполнена операция сегментации.

В основе сегментации лежат алгоритмы Ветерби и юниграмм. Ключевой идеей алгоритма Витерби является поэтапное сравнение всех кодовых слов с наблюдением и отбрасывании тех из них, которые находятся на большем расстоянии от данного наблюдения, чем некоторые другие пути решетчатой диаграммы. Более подробно процедура декодирования может быть описана следующим образом.

Пусть i-ым шагом декодирования будет временной интервал, в течение которого принимается i-я n-символьная кодовая группа наблюдения. Перед этим моментом рассматриваемые пути могут проходить через один из 2^{m-1} узлов решетчатой диаграммы. Рассмотрим i-ый шаг выполнения алгоритма.

1. Определяется хэммингово расстояние между каждой из ветвей решетчатой диаграммы и принятой n-символьной кодовой группой. Так как из каждого из 2^{m-1} узлов выходят две ветви, всего вычисляется 2^m расстояний.
2. Рассматриваются две ветви, которые идут из разных предшествующих состояний к каждому из 2^{m-1} узлов.

2.1. Для того чтобы получить новые расстояния, хэмминговые расстояния, отвечающие данным ветвям, прибавляется к накопленному до i-го шага расстоянию Хэмминга двух соответствующих путей. В результате чего будет получено накапливаемое расстояние пути, называемое метрикой. Ниже используется терминология из работы.

2.2. Идет сравнение метрик двух разных путей, идущих в одно и то же состояние. В результате чего, путь, который окажется на большем наблюдения, чем другой, будет отброшен и больше не будет учитываться. Путь, который останется, называется выжившим путем.

3. Для всех 2^{m-1} выживших путей запоминаются значения их метрик и декодер готов к переходу на $(i+1)$-й шаг процедуры.

В результате выполнения данных операций можно сделать вывод о том, что ресурсосбережением алгоритма можно определить, как отбрасывание на каждом этапе ровно половины из 2^m возможных путей, ведущих в 2^{m-1} узлов решетки. После чего, число выживших путей остается постоянным и равным 2^{m-1} различных состояний вне зависимости от величины соревнующихся кодовых слов, число которых удваивается на каждом шаге алгоритма декодирования.

Для нахождения набора n-грамм слов документа используются юниграммы. Они представляют собой самую простую модель для подхода n-грамм, состоящую из всех отдельных слов, присутствующих в тексте.

6. Алгоритмы авторской программы

Рассмотрим алгоритм одной итерации работы авторской программы получения уровней для одного исходного открытого текста с выбранным уровнем зашумления, а также алгоритм получения уровней для множества текстов одной длины с различными уровнями зашумления.

6.1. Получение уровней для одного исходного открытого текста

Алгоритм выполняется над одним исходным текстом $t \in T$ длины L из коллекции T с фиксированным относительным уровнем зашумления $n_{\%}$. Текст $t \in M$, $M \subseteq X$, $X = I^L$, где M – множество содержательных текстов из множества открытых текстов X длиной L алфавита I.

- Выделение ключевых слов из открытого неискаженного текста. Используя алгоритм TF-IDF над открытым исходным текстом $t \in T$, определяется вес каждого слова $tfidf(w_i)$, $w_i \in t$. Далее выбирается набор ключевых слов $W_t = \{w_i | tfidf(w_i) > T_{min}\}$. Вес слов в наборе W_t не учитывается, все слова считаются одинаково значимыми.
- Искажение (зашумление) исходного открытого текста. На вход подпрограмме зашумления подается текст t и выбранный $n_{\%}$ Результатом искажения является зашумленный текст t'.
- Исправление орфографических ошибок. На вход подпрограмме исправления орфографических ошибок подается зашумленный текст t'. С помощью Yandex.Speller обнаруживаются и исправляются орфографические ошибки. В результате получаем искаженный и орфографически исправленный текст t''.
- Разделение объединенных в результате зашумления пробелов слов. Запускается подпрограмма сегментации, на вход ей подается t''. В результате входной текст преобразуется в текст t''' (орфографически исправленный зашумленный текст с сегментированными словами, называемый восстановленным текстом).
- Выделение ключевых слов из восстановленного текста. Используя алгоритм TF-IDF над восстановленным текстом t''', определяется вес каждого слова $tfidf(w_i''')$, $w_i''' \in t'''$. Далее выбирается набор ключевых слов $W_{t'''} = \{w_i''' | tfidf(w_i''') > T_{min}\}$. Вес слов в наборе $W_{t'''}$ не учитывается, все слова считаются одинаково значимыми.
- Сравнение ключевых слов. Проводится алгоритм сравнения ключевых слов. Сравниваются множества W_t и $W_{t'''}$ ключевых слов из текстов t и t' соответственно. Находится $W_t \cup W_{t'''}$ общих ключевых слов. Далее определяется уровень содержания восстановленного текста

$$CL\left(W_t, W_{t'''}\right) = \frac{\left|W_t \bigcup W_{t'''}\right|}{\left|W_t\right|}$$ (CL = content level).

Например, если из исходного текста выделены ключевые слова {«война», «победа», «план»} (множество из трех элементов), а из восстановленного текста {«война», «план», «папка», «карта» }, то УС составит $2/3 \approx 66\%$, как отношение числа элементов в пересечении множеств ({«война», «план» }) к числу элементов в множестве ключевых слов исходного текста ({«война», «победа», «план» }). Полученный показатель $CL\left(W_t, W_{t'''}\right)$ может быть сведен к одному из уровней: «почти все понятно» ($66-100\%$), «средне понятно» ($33-66\%$), «плохо понятно» ($0-33\%$).

6.2. Получение уровней для множества текстов одинаковой длины

Для анализа зависимости уровней содержания восстановленных текстов CL от длины L и уровня зашумления $n_{\%}$ необходимо провести серию повторений выполнения программы для разных входных параметров.

При повторении алгоритма исходные тексты t разделяются на группы по длинам L и для каждого из уровней зашумления $n_{\%1}, n_{\%2}, \ldots, n_{\%k}$ для каждого из текстов многократно определяется CL. В результате получаются тройки $h_j = \left(L_j, \left(n_{\%}\right)_j, CL_j\right)$.

Далее для каждого из $L, n_{\%}$ определяется среднее значение $CL_{avg}\left(L, n_{\%}\right)$ уровней содержания восстановленного текста CL_j:

$$H_{L,n_{\%}} = \left\{h_j : L_j = L, \left(n_{\%}\right)_j = n_{\%}\right\},$$

$$CL_{avg}\left(L, n_{\%}\right) = \frac{\sum_{H_{L,n_{\%}}} CL_j}{\left|H_{L,n_{\%}}\right|}$$

Параметр $CL_{avg}\left(L, n_{\%}\right)$ будем называть среднетекстовый уровень содержания восстановленного текста.

7. Полученные результаты

Среднетекстовые уровни содержания восстановленного текста, полученные в результате повторения выполнения программы для $L \in \{100, 200, 300, 400\}$ и $n_{\%} \in \{5\%, 10\%, 15\%, 20\%, 25\%, 30\%, 35\%, 40\%, 45\%, 50\%\}$ представлены в табл. 1. Цветом фона и толщиной границ показаны уровни: «плохо понятно» (серый цвет фона), «средне понятно» (толстые границы), «почти все понятно» (без выделения).

Таблица 1.

Среднетекстовые уровни для текстов длин 100-400 символов

		L (длина текста, символов)			
		100	200	300	400
n% (уровень зашумления, %)	5	100	100	100	100
	10	96,5	98,5	100	99
	15	79	87	91	61,5
	20	57,5	50	39	8,5
	25	39	19,5	7,5	0,5
	30	20	6	1,5	0
	35	5	1,5	0	0
	40	3,5	0,5	0	0
	45	0	0	0	0
	50	0,5	0	0	0

Ниже представлен график, отражающий зависимость полученных среднетекстовых уровней содержания от уровня зашумления для разных длин текстов. Вертикальная ось – среднетекстовый уровень содержания (%), горизонтальная ось – уровень зашумления (%). Линии с маркерами последовательно соединяют уровни для текстов одной длины: маркеры-круги – 100 символов, маркеры-кресты – 200 символов, маркеры-ромбы – 300 символов, маркеры-треугольники – 400 символов. Горизонтальными прямыми показаны границы уровней: 33% – между «плохо понятно» и «средне понятно», 66% – между «средне понятно» и «почти все понятно».

УДК 003.26.09 DOI: 10.21681/2311-3456-2019-5-42-50

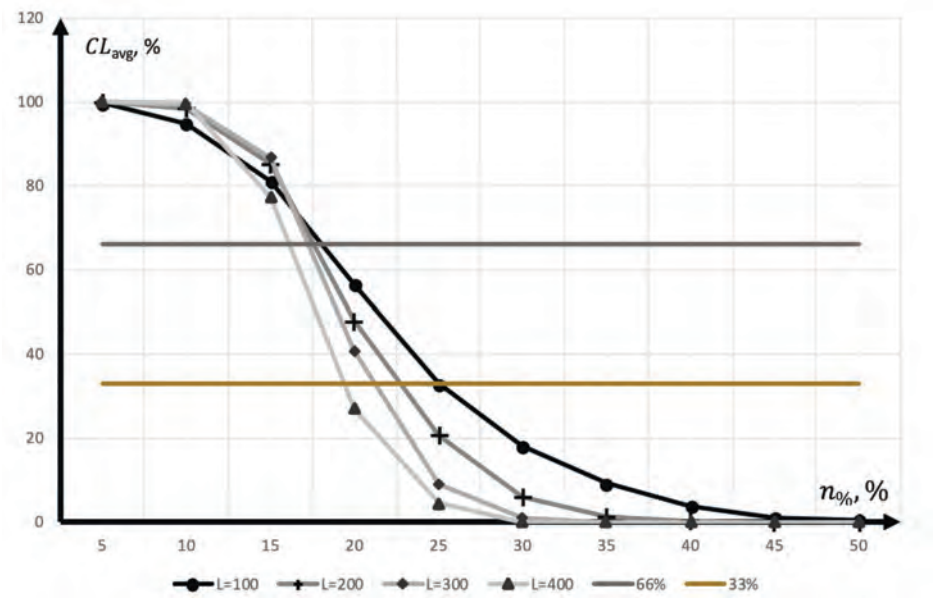

Рис. 1. График зависимости для текстов длин 100-400 символов

Также программа была выполнена для $L \in \{1000, 1500, 2000\}$ и $n_{\%} \in \{5\%, 10\%, 15\%, 20\%, 25\%, 30\%, 35\%, 40\%, 45\%, 50\%\}$. График зависимости приведен ниже. Вертикальная ось – среднетекстовый уровень содержания (%), горизонтальная ось – уровень зашумления (%). Линии с маркерами последовательно соединяют уровни для текстов одной длины: маркеры-кресты – 1000 символов, маркеры-ромбы – 1500 символов, маркеры-треугольники – 2000 символов. Горизонтальными прямыми показаны границы уровней: 33% – между «плохо понятно» и «средне понятно», 66% – между «средне понятно» и «почти все понятно».

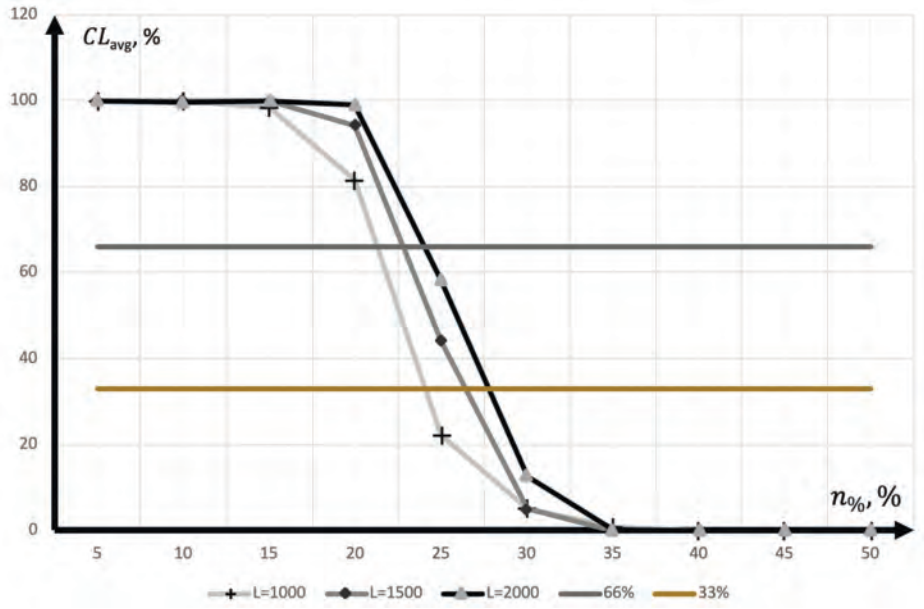

Рис. 2. График зависимости для текстов длин 1000-2000 символов

8. Выводы

Исходя из полученных результатов, можно сделать следующие выводы:

1. При относительном уровне зашумления 20% тексты малых длин (100-400 символов) становятся нечитаемыми, нелогичными и непонятными («плохо понятно» по введенной шкале).
2. Для текстов длин 1000-2000 символов критический относительный уровень зашумления составляет 25%, именно при таком значении $n_{\%}$ невозможно установить смысл текста.
3. Для текстов длин 1000-2000 символов происходит резкий спад читаемости текста после относительного уровня зашумления в 20%, для текстов малой длины такого эффекта не наблюдается: спад среднетекстовых уровней содержания можно назвать постепенным с нарастанием относительного уровня зашумления.

Литература

1. С. В. Запечников, О.В. Казарин, А.А. Тарасов. Криптографические методы защиты информации. М., Юрайт, 2017.
2. Rubinstein-Salzedo S. The Vigenere Cipher / S. Rubinstein-Salzedo // Cryptography. 2018. №4. p. 41-54.
3. Бабаш А.В. Обобщенная модель шифра / А.В. Бабаш // Интеллектуальные системы в информационном противоборстве. 2015. №1. С.9-14.
4. Aized Amin Soofi. An Enhanced Vigenere Cipher For Data Security / Aized Amin Soofi [and etc.] // International journal of Scientific & Technology research. 2016. №3. p.141 – 145.
5. Abiodun Esther Omolara. An Enhanced Practical Difficulty of One-Time Pad Algorithm Resolving the Key Management and Distribution Problem / Abiodun Esther Omolara [and etc.] // Proceedings of the International MultiConference of Engineers and Computer Scientists 2018. 2018. №1. p.1 – 7.
6. Jean-Philippe Aumasson. Serious Cryptography. A Practical Introduction to Modern Encryption: practical guide / Jean-Philippe Aumasson. – San Francisco: no starch press, 2017. 312 p.
7. Михайлов, Д.В. Выделение знаний и языковых форм их выражения на множестве тематических текстов: подход на основе меры TF-IDF / Д.В. Михайлов, А.П. Козлов, Г.М. Емельянов // Компьютерная оптика. 2015. Т. 39, № 3. С. 429-438.
8. Воронцов, К.В. Многокритериальные и многомодальные вероятностные тематические модели коллекций текстовых документов / К.В. Воронцов, А.А. Потапенко, А.И. Фрей, М.А. Апишев, Н.В. Дойков, А.В. Шапу-лин, Н.А. Чиркова // 10-я Междунар. конф. ИОИ-2014: Тезисы докладов. 2014. С. 198.
9. Осипова Ю.А. Применение кластерного анализа методом k-средних для классификации текстов научной направленности // Математические структуры и моделирование. 2017. №3 (43). С. 108-121.
10. Яндекс.Спеллер: Документация – О сервисе [Электронный ресурс] URL: https://tech.yandex.ru/speller/doc/dg/concepts/speller-overview-docpage/ (дата обращения: 04.06.2019 г.)
11. Лавошникова Э.К. «Проблемные» слова как причина пропуска ошибок при компьютерной проверке орфографии // Текст. Книга. Книгоиздание. 2017. № 15. С. 104–112.
12. Нгуен Ч.Т. Алгоритм распознавание зашумленных речевых команд на основе пробных спектральных преобразований входного сигнала // Известия ТулГУ. Технические науки. 2014. №5. С.236-232.
13. Ковыршин И.О. Параметры для определения зашумленных речевых сообщений / Ковыршин И.О., Скурнович А. В. // Новые информационные технологии в автоматизированных системах. 2016. №19. С. 127-130.
14. Ярмоленко А.С., Писецкая О.Н., Кудаева О.А. Система распознавания площадных объектов с зашумленными обучающими образами. // Вестник Белорусской гос.с.-х. акад. – 2015. №2. С. 129.
15. Корсунов Н.И. Метод обучения персептрона распознаванию текстовых символов при зашумлениях / Н. И. Корсунов, К. В. Лысых, Д. А. Торопчин // Научные ведомости БелГУ. Сер. Экономика. Информатика. – 2015. – №13(210), вып.35/1.-С. 99-103.
16. Самойлин Е.А. Программная модель для исследования эффективности процедур выделения контуров зашумленных изображений / Самойлин Е. А. Карпов С. А. // Программные продукты и системы. 2018. №4. С. 734-739.
17. Ю. В. Косолапов, О. Ю. Турченко, «Поиск информационного сообщения в зашумлённых кодовых блоках при многократной передаче данных», ПДМ. Приложение, 2016. № 9. С. 55–57.
18. Ferrand A. Using the NoiSee workflow to measure signal-to-noise ratios of confocal microscopes / A. Ferrand [and etc.] // Scientific Report. 2019. №1165. p.1-28.
19. Толстунов В.А. Нелинейный сглаживающий фильтр с показательно-степенными весами / В.А. Толстунов // Технические науки. 2015. №2(15). С.10-18.
20. Feng X. Reconstruction of noisy images via stochastic resonance in nematic liquid crystals / X. Feng [and etc.] // Scientific Reports. 2019. №3976. p.1 – 15.
21. Крашенинников В.Р. Зашумление эталонов в задачах обнаружения и распознавания сигналов на фоне помех / В.Р. Крашенинников, А.И. Армер // Вестник УлГТУ. 2004. №2. С.54-57.
22. Koohian A. Joint channel and phase noise estimation for mmWave full-duplex communication systems / A. Koohian [and etc.] // Eurasip Journal on Advances in Signal Processing. 2019. №18. p.1 – 12.
23. Shaaban R. Visible light communication security vulnerabilities in multiuser network: power distribution and signal to noise ratio analysis / R. Shaaban [and etc.] // Lecture Notes in Networks and Systems. 2019. №2020. p.1 – 13.

24. Lira de Queiroz W.J. Signal-to-noise ratio estimation for M-QAM signals in η−μand κ−μfading channels / W.J. Lira de Queiroz [and etc.] // Eurasip Journal on Advances in Signal Processing. 2019. №20. p.1 – 17.
25. Roy S. Fundamental noisy multiparameter quantum bounds / S. Roy // Scientific Reports. 2019. №1038. p.1 – 15.
26. Васильева И.Н. Криптографические методы защиты информации: учебное пособие / И.Н. Васильева. М.: Юрайт, 2016. – 64 с.

SPREADING BORDERS OF VIGENERE CIPHER DECRYPTION METHODS

Babash A.V.[6], Guzovs R.[7], Kasatkin S.V.[8], Prohorov A.N.[9], Slimov N.A.[10]

Purpose: *to introduce strict formalized model of plaintext context, noisy text, to determine the lowest limit of the plaintext noisiness when the content cannot be understood by a native speaker.*

Research methods: *a custom Vigenere cipher decryption which uses as key substandard gamma representing periodical noisy sequence.*

Results: *the formalization of the task of decrypting the cipher of random gamming, the Vigenere cipher, the concept of the content of the distorted plaintext are given. The algorithms of extracting plaintext content (keywords), its suppression, recovery of the noisy text content with spelling correction and segmentation have been developed. The scale of text understanding levels is introduced. The experimental value of the average text level of content understanding of the restored noisy texts with the length of 100 up to 2000 characters have been obtained based on the repeated random noisiness of the similar subjects texts.*

Keywords*: Vigenere cipher, plaintext, plaintext content, noisy text; informative text.*

References

1. S. V. Zapechnikov, O.V. Kazarin, A.A. Tarasov. Kriptograficheskie metody zashchity informacii. M., YUrajt, 2017.
2. Rubinstein-Salzedo S. The Vigenere Cipher / S. Rubinstein-Salzedo // Cryptography.2018. №4. p. 41-54.
3. Babash A.V. Generalized cipher model / A.V. Babash // Intellectual systems in the information confrontation. 2015. №1. p.9-14.
4. Aized Amin Soofi. An Enhanced Vigenere Cipher For Data Security / Aized Amin Soofi [and etc.] // International journal of Scientific & Technology research. 2016. №3. p.141 – 145.
5. Abiodun Esther Omolara. An Enhanced Practical Difficulty of One-Time Pad Algorithm Resolving the Key Management and Distribution Problem / Abiodun Esther Omolara [and etc.] // Proceedings of the International MultiConference of Engineers and Computer Scientists 2018. 2018. №1. p.1–7.
6. Jean-Philippe Aumasson. Serious Cryptography. A Practical Introduction to Modern Encryption: practical guide / Jean-Philippe Aumasson. San Francisco: no starch press, 2017. 312p.
7. Mihaylov, D.V. Highlighting knowledge and linguistic forms of their expression on a variety of thematic texts: an approach based on the TF-IDF measure / D.V. Mihaylov, A.P. Kozlov, G.M. Emelyanov // Computer optics. 2015. C. 39, № 3. P. 429-438.
8. Vorontsov, K.V. Multi-criteria and multimodal probabilistic thematic models of collections of text documents / K.V. Vorontsov, A.A. Potapenko, A.I. Frey, M.A. Apishev, N.V. Doykov, A.V. Shapu-ling, N.A. Chirkova // 10th Intern. conf. IOI-2014: Abstracts. 2014. p. 198.
9. Osipova Yu.A. The use of cluster analysis by the k-means method for the classification of scientific texts // Mathematical structures and modeling. 2017. No.3 (43). Pp. 108-121.
10. Yandex.Speller: Documentation – About the service [Electronic resource] URL: https://tech.yandex.ru/speller/doc/dg/concepts/speller-overview-docpage/.
11. Lavoshnikova E.K. «Problem» words as a reason for missing errors in computer spell check. Text. Book. Book publishing. 2017. № № 15. Pp. 104–112.
12. Nguyen H.T. Algorithm for the recognition of noisy speech commands based on trial spectral transformations of the input signal // Bulletin of TSU. Technical science. 2014. №5. Pp.236-232.
13. Kovyrshin I.O. Parameters for the definition of noisy voice messages / Kovyrshin I.O., Skurnnovich A.V. // New information technologies in automated systems. 2016. №19. P. 127-130.

6 Alexander Babash, Dr.Sc., Professor, National Research University Higher School of Economics, Moscow, Russia. E-mail: ababash@hse.ru
7 Rihard Guzovs, student, Master's Programme National Research University Higher School of Economics, Moscow, Russia.
8 Semyon Kasatkin, student, Master's Programme National Research University Higher School of Economics, Moscow, Russia.
9 Arseny Prohorov, student, Master's Programme National Research University Higher School of Economics, Moscow, Russia.
10 Nikita Slimov, student, Master's Programme National Research University Higher School of Economics, Moscow, Russia.

14. Yarmolenko A.S., Pisetskaya ON, Kudaeva O.A. The recognition system of polygon objects with noisy training images. // Bulletin of the Belarusian state.s.-kh. Acad. 2015. №2. Pp. 129.
15. Korsunov N.I. The perceptron's method of learning the recognition of text characters in noise / N. I. Korsunov, K. V. Lysykh, D. A. Toropchin // Scientific Gazette of BelSU. Ser. Economy. Computer science. 2015. №13 (210), vol.35 / 1. Pp. 99-103.
16. Samoylin E.A. Software model for the study of the effectiveness of procedures for the selection of contours of noisy images / Samolin E. A. Karpov S. A. // Software products and systems. 2018. №4. Pp. 734-739.
17. Yu. V. Kosolapov, O. Yu. Turchenko, «Search for an informational message in noisy code blocks with multiple data transmissions», Prikl. Appendix, 2016. No. 9. P. 55–57.
18. Ferrand A. Using the NoiSee workflow to measure signal-to-noise ratios of confocal microscopes / A. Ferrand [and etc.] // Scientific Report. 2019. №1165. Pp.1-28.
19. Tolstunov V.A. Nonlinear smoothing filter with exponential power scales / V.A. Tolstunov // Technical Sciences. 2015. №2 (15). Pp.10-18.
20. Feng X. Reconstruction of noisy images via stochastic resonance in nematic liquid crystals / X. Feng [and etc.] // Scientific Reports. 2019. №3976. Pp.1 – 15.
21. Krasheninnikov V.R. Pattern noise in the signal detection and recognition tasks with interference / V.R. Krasheninnikov, A.I. Armer // Vestnik of UlSTU. 2004. №2. Pp.54-57.
22. Koohian A. Joint channel and phase noise estimation for mmWave full-duplex communication systems / A. Koohian [and etc.] // Eurasip Journal on Advances in Signal Processing. 2019. №18. Pp.1 – 12.
23. Shaaban R. Visible light communication security vulnerabilities in multiuser network: power distribution and signal to noise ratio analysis / R. Shaaban [and etc.] // Lecture Notes in Networks and Systems. 2019. №2020. Pp.1 – 13.
24. Lira de Queiroz W.J. Signal-to-noise ratio estimation for M-QAM signals in η−μand κ−μfading channels / W.J. Lira de Queiroz [and etc.] // Eurasip Journal on Advances in Signal Processing. 2019. №20. Pp.1 – 17.
25. Roy S. Fundamental noisy multiparameter quantum bounds / S. Roy // Scientific Reports. 2019. №1038. Pp.1 – 15.
26. Vasilyeva I.N. Information protecting cryptographic methods: a tutorial / I.N. Vasilieva. M.: Yurayt, 2016. 64 p.