УДК 519.722

А.Г. Малашина

Модификация одного алгоритма восстановления текстовых сообщений и математическая модель распределения числа осмысленных текстов

Предложена модификация одного из методов восстановления зашифрованного текстового сообщения в поточных шифрах. Сгенерированы словари *n*-грамм, используемые для реализации данного метода, и исследованы их статистические свойства. Предложена методика оценки полноты и достаточности составляемых словарей. Построена математическая модель распределения осмысленных текстов на русском языке, возникающая при восстановлении отдельных участков зашифрованного текста. Найдены оптимальные параметры для предлагаемого алгоритма.

Ключевые слова: словари *n*-грамм, энтропия *n*-грамм, защита информации, информационная безопасность, осмысленные тексты.

В ряде задач по анализу алгоритмов защиты информации возникает ситуация, когда относительно знаков неизвестного исходного текстового сообщения появляется информация о возможных вариантах его знаков [1]. Если количество вариантов для каждого знака относительно невелико, то исходное сообщение достаточно просто восстанавливается полностью. В противном случае если количество вариантов знаков можно считать случайным, появляется возможность построить процедуру поиска и последующего восстановления отдельных участков сообщения, на которых сосредоточены знаки с относительно небольшим количеством возможных вариантов.

Модификация алгоритма восстановления текстового сообщения

В традиционном подходе для каждого символа шифрованного сообщения известно ограниченное число вариантов знаков исходного текста. В этом случае если под каждым неизвестным знаком сообщения выписать все известные для него варианты, посредством перебора всех возможных комбинаций вариантов символов удаётся восстановить осмысленный исходный текст (рис. 1) [1].

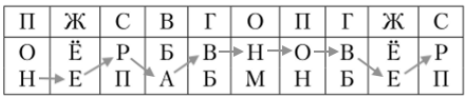

Рис. 1. Классическая схема бесключевого восстановления исходного текста

Однако с увеличением количества вариантов знаков для каждого символа сообщения восстановление исходного текста подобным образом становится затруднительным из-за высокой степени неоднозначности выбора осмысленного текста, возникающей в процессе отыскания подходящих вариантов восстановления. Для русского языка однозначное восстановление теоретически возможно, когда количество возможных вариантов символов исходного текста не превышает 16 знаков [1]. Кроме того, на практике известная информация о символах исходного текста распределена неравномерно. В этом случае представляется возможным восстановить исходный текст лишь на отдельных участках шифрованного сообщения, т.е. там, где сосредоточены знаки с относительно небольшим количеством возможных вариантов.

В данной статье предлагается алгоритм восстановления отдельных отрезков сообщения, основанный на использовании словарей *n*-грамм [3]. Словари составляются на основе собранного языкового корпуса – общественно-политических текстов на русском языке [2].

Пусть для каждого символа шифрованного сообщения удалось построить множество вариантов его знака, среди которых находится истинный символ. Число таких возможных вариантов может иметь различное вероятностное распределение. Восстановление неизвестного сообщения заключается в нахождении осмысленного текста среди всех построенных комбинаций текста из известных вариантов символов исходного текста.

Для того чтобы найти участки сообщения, для которых число подходящих вариантов дешифрования близко к единице, предлагается оценка L_i: *среднее геометрическое значение* количества возможных вариантов для соответствующего отрезка. Отбираются только те отрезки сообщения, для которых L_i *не превышает* заранее заданной критической границы $L_{\text{кр}}$:

$$L_i = \sqrt[n]{l_{i_1} \cdot l_{i_2} \cdot \ldots \cdot l_{i_n}} \le L_{\text{кр}}, \qquad (1)$$

где n – длина отрезка, l_{i_j} – количество вариантов символа для j-го знака в i-й n-грамме.

Процедура восстановления отобранных участков сообщения заключается в переборе всех возможных комбинаций, которые можно построить, используя известные варианты символов для каждого знака выбранного отрезка. Если построенный вариант является словарной величиной, то считается, что он представляет собой часть осмысленного текста, и рассматривается как потенциальный отрезок исходного текста сообщения. Если же состав-

ленная *n*-грамма не присутствует в словаре, то предполагается, что она является текстом случайной структуры и отвергается как ложный вариант дешифрования.

В зависимости от количества найденных осмысленных вариантов отрезок сообщения считается успешно восстановленным либо невосстановленным. Предлагается следующая оценка максимально допустимого числа вариантов дешифрования *N* для участков сообщения заданной длины:

$$N = 2^{0,1n}, \qquad (2)$$

где *n* – длина участка сообщения.

Статистические свойства словарей различных языковых моделей

Исследование возможности восстановления отдельных участков сообщения в описанных выше ситуациях основывается на создаваемых словарях, поэтому целесообразно изучение их статистических свойств [2]. В данном разделе приведены результаты исследования вероятностного и статистического распределения слов и *n*-грамм русского языка.

Для составления словарей используются две языковые модели: лексическая (единица анализа – токены) и *n*-граммная (*n* символов). Словари генерируются из текстового корпуса объёмом 10 млн символов (табл. 1).

Таблица 1

Объём составленных словарей

Единицы анализа	Объём словаря
Токены (слововхождения)	163 091
10-граммы	6 217 191
15-граммы	9 482 897
20-граммы	10 372 296
25-граммы	10 629 589

Теоретический подход к оценке величины покрытия

Чтобы использовать созданные словари в алгоритме восстановления участков сообщения, они должны быть достаточно полными. Полнота словарей обусловливается их покрытием.

Предлагается следующий подход к теоретической оценке покрытия словарей.

Пусть получен словарь объёмом *N* элементов. Поскольку он составляется на материале ограниченного объёма, данный словарь не обладает полным покрытием. Для оценки размера словаря, компенсирующего недостаток покрытия исходного, предполагается использовать следующую теоретическую оценку:

$$\tilde{N} = \frac{N}{1 - n/N}, \qquad (3)$$

где *N* – размер исходного словаря, *n* – число однократно встречающихся элементов словаря.

Тогда $1 - n/N$ – доля повторяющихся элементов в словаре, а $(1 - \frac{n}{N})N$ – число повторяющихся элементов исходного словаря. Очевидно, $(1 - n/N)N \le N$.

На основе экспериментальных данных известно, что до определённого момента число новых слов в словаре растёт с линейной скоростью [2]. Чтобы получить словарь, состоящий из *N* повторяющихся элементов, нужно увеличить его объём в $1/(1 - n/N)$ раз.

Таким образом, новый словарь объёмом $\tilde{N} = \frac{N}{1 - \frac{n}{N}}$ содержит около *N* повторяющихся элементов. Тогда покрытие нового словаря исходным оценивается как

$$\frac{\tilde{N}}{N} \cdot 100\% = (1 - \frac{n}{N}) \cdot 100\%. \qquad (4)$$

Следовательно, теоретической оценкой покрытия словаря можно считать долю повторяющихся элементов $1 - n/N$ в словаре объемом *N*.

Величина $\tilde{N}$ гипотетического словаря используется в дальнейшем для оценки объёма словаря на следующей итерации (после очередного увеличения длины исходного материала) и сравнения ожидаемой величины покрытия с реальной.

Покрытие словарей может быть найдено также эмпирическим методом (табл. 2). Для этого экспериментально оценивается доля тестового экземпляра текста, покрываемая выбранным словарём.

Таблица 2

Величина покрытия словарей

Словарь	Теоретическое покрытие, %	Эмпирическое покрытие, %
Токенов (слововхождений)	70,63	76,45
10-грамм	19,95	39,71
15-грамм	7,21	12,03
20-грамм	3,27	3,12
25-грамм	2,07	0,84

Информационная энтропия и закон Ципфа

Многие языки имеют распределение слов, близкое к распределению Парето, т.е.

$$p_i = \frac{1}{c \cdot i}, \quad c = \sum_{i=1}^{m^n} \frac{1}{i} = \text{const}, \qquad (5)$$

где p_i – вероятность появления *i*-го слова, *m* – размер алфавита, *n* – длина слова [4].

Таким образом, частота использования слов, упорядоченных по убыванию их встречаемости, обратно пропорциональна их порядковому рангу. Данное соотношение называется законом Ципфа [5].

Для проверки качества и естественности созданных словарей проводится проверка их соответствия закону Ципфа. Экспериментальные данные сопоставляются с идеальными значениями распределения Парето (рис. 2).

Основополагающим принципом рассматриваемого алгоритма является предположение о том, что *n*-грамма осмыслена только тогда, когда она присутствует в словаре. Поэтому объём *D* соответствующего словаря *n*-грамм – это эмпирическая оценка коли-

чества осмысленных текстов в русском языке длины n. Тогда верно соотношение

$$D = 2^{H \cdot n}, \qquad (6)$$

где n – длина текста, H – энтропия источника [1].

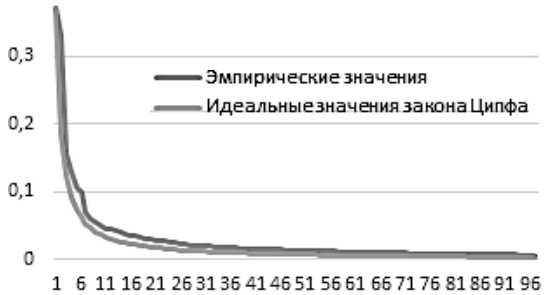

Рис. 2. Закон Ципфа

В таком случае энтропию n-грамм можно оценить, опираясь на эмпирическую величину D:

$$H_n = \log_2 \frac{D}{n}. \qquad (7)$$

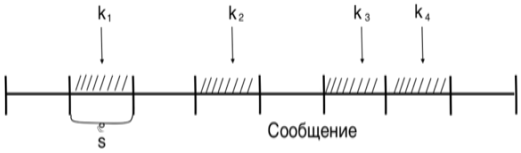

Рис. 3. Распределение числа осмысленных текстов

В табл. 3 приведены значения информационной энтропии n-грамм исследуемого текстового корпуса: N_{10^6} – словарь на материале 1 млн символов, $\tilde{N}_{10^6}$ – теоретический словарь, компенсирующий неполноту словаря N_{10^6} (теоретическая оценка размера N_{10^7}), N_{10^7} – словарь на материале 10 млн символов.

Таблица 3

Значения энтропии *n*-грамм

n	N_{10^6}	$\tilde{N}_{10^6}$	N_{10^7}
10	1,94	2,20	2,26
15	1,32	1,57	1,55
20	0,99	1,22	1,17
25	0,79	0,99	0,93

Значения энтропии n-грамм, полученные на основе теоретической оценки размера исходного словаря, близки к значениям энтропии, полученным эмпирически после увеличения исходного текстового материала в 10 раз. Это говорит о том, что предлагаемая формула (3) позволяет получить правдоподобную оценку размера полного словаря n-грамм.

С ростом величины n энтропия n-грамм приближается к реальному значению энтропии русского языка.

Математическая модель распределения числа осмысленных текстов

Количество осмысленных текстов и вероятность их появления при дешифровании участков сообщения могут быть представлены в виде математической модели.

Пусть $N = m^s$ – общее количество s-грамм, которые могут быть построены из символов алфавита мощностью m, D – число осмысленных текстов длины s в том же алфавите, n – число построенных вариантов для восстанавливаемого отрезка сообщения (в рассматриваемой модели – разное для каждого дешифруемого отрезка из-за различных длин колонок), k – количество найденных осмысленных текстов для восстанавливаемого отрезка сообщения.

Тогда вероятность того, что в выборке из n различных вариантов восстановления, полученных из множества всех возможных вариантов N, содержащего D осмысленных текстов, ровно k вариантов окажутся осмысленными, описывается гипергеометрическим распределением, т.е.

$$P(k) = \frac{C_D^k C_{N-D}^{n-k}}{C_N^n}. \qquad (8)$$

Если восстановлению подвергаются участки неизвестного текста длиной s символов с использованием словаря s-грамм объёма D, то вероятность появления k осмысленных вариантов восстановления описывается табл. 4 (l_i – количество вариантов символов для i-го знака, объём алфавита – 35 символов).

Таблица 4

Гипергеометрическая модель

Число вариантов	Найденных	Не найденных	Всего
Осмысленных	k	D-k	D
Неосмысленных	n-k	N-D-n+k	N-D
Всего	$n = l_1 l_2 \cdot \ldots \cdot l_s$	N-n	$N = 35^s$

В теории и на практике общее число текстов N значительно превышает размер выборки n ($n << N$). В этом случае гипергеометрическое распределение хорошо аппроксимируется биномиальным распределением с параметрами n и D/N.

Для сопоставления теоретических вероятностей с эмпирическими проводится экспериментальное восстановление 5-грамм тестового сообщения. Варианты возможных символов для каждого знака сообщения генерируются случайно на основе дискретного равномерного распределения. Для восстановления выбираются отрезки со средней геометрической глубиной колонок не больше 12 символов. Для каждого восстанавливаемого отрезка фиксируется общее число всех возможных вариантов и количество осмысленных (т.е. присутствующих в словаре).

На рис. 4 представлены теоретические и эмпирические вероятности появления различного числа осмысленных текстов при восстановлении 5-грамм сообщения. Вероятность появления k осмысленных вариантов восстановления с ростом k резко уменьшается.

В гипергеометрической модели наиболее вероятные значения располагаются при k от 50 до 70. В эмпирической выборке наиболее вероятные значе-

ния k лежат в интервале от 40 до 75, что полностью оправдывает теоретические ожидания.

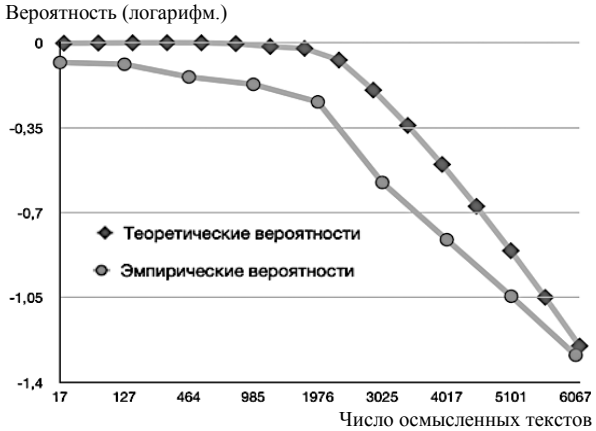

Рис. 4. Распределение числа осмысленных текстов

Такая степень неоднозначности слишком высока для восстановления истинного исходного текста, поэтому на практике дешифрование 5-грамм вышеописанным способом практически невозможно.

Используя гипергеометрическое моделирование, определяется наиболее вероятное количество осмысленных текстов M_o (мода), которое может появиться при дешифровании s-грамм различной длины (с учётом того, что истинный осмысленный вариант всегда присутствует в выборке). $\bar{X}$ – выборочное среднее найденных осмысленных вариантов дешифрования (табл. 5).

Таблица 5

Вероятное число осмысленных текстов

Длина отрезка	M_o (теор.)	$\bar{X}$ (практ.)
10	8–12	16,38
15	3	2,64
20	1	1

С увеличением длины отрезка сообщения уменьшается степень неоднозначности восстановления исходного текста. Проведённое моделирование позволяет выбрать наиболее оптимальный параметр алгоритма – длину восстанавливаемого отрезка сообщения – на основе максимально эффективного соотношения между степенью неоднозначности восстановления исходного текста и покрытием соответствующего словаря. Для эффективной работы рассматриваемого алгоритма предлагается использовать s-граммы длиной от 15 до 20 символов.

Заключение

В данной работе был рассмотрен алгоритм восстановления отдельных участков неизвестного исходного сообщения при заданных вариантах его знаков. Для реализации данной процедуры созданы специальные словари и исследованы их статистические свойства. Кроме того, разработана методика теоретической оценки покрытия словарей и проведена экспериментальная оценка её релевантности. На основе гипергеометрического распределения смоделированы вероятности появления различного количества осмысленных текстов, что позволило определить оптимальный параметр предлагаемого алгоритма: восстановление отрезков длиной от 15 до 20 символов.

Литература

1. Алферов А.П. Основы криптографии: учеб. пособие. – 3-е изд., испр. и доп. / А.П. Алферов, А.Ю. Зубов, А.С. Кузьмин, А.В. Черемушкин. – М.: Гелиос АРВ, 2005. – 408 с.

2. Малашина А.Г. Статистический анализ языковых моделей русского языка на основе новостного текстового корпуса // Межвуз. науч.-техн. конф. студентов, аспирантов и молодых специалистов им. Е.В. Арменского: матер. конф. – 2020 (в печати).

3. Малашина А.Г. Алгоритм восстановления отдельных частей сообщения по информации о возможных значениях его знаков // Межвуз. науч.-техн. конф. студентов, аспирантов и молодых специалистов им. Е.В. Арменского: матер. конф. – 2019. – С. 215–217.

4. Kechedzhy K.E. Rank distributions of words in additive many-step Markov chaons and the Zipf law / K.E. Kechedzhy, O.V. Usatenko, V.A. Yampol'skii // Phys. Rev. E. – 2005. – Vol. 72.

5. Gelbukh A. Zipf and Heaps Laws' Coefficients Depend on Language / A. Gelbukh, G. Sidorov. – Mexico City: Conference on Intelligent Text Processing and Computational Linguistics, 2001.

Малашина Анастасия Геннадьевна
Аспирант каф. компьютерной безопасности НИУ ВШЭ
Таллинская ул., 34, г. Москва, Россия, 123592
ORCID: 0000-0002-5163-0593
Тел.: +7 (495-7) 72-95-90, доп. 15125
Эл. почта: amalashina@hse.ru