Ежегодная международная научно-практическая конференция
«РусКрипто’2021»

Алгоритм восстановления отдельных частей текстовых сообщений по информации о возможных вариантах его знаков

Малашина Анастасия,
Аспирант, НИУ «Высшая школа экономики»

Восстановление текстового сообщения

Возможно восстановить отдельные участки зашифрованного сообщения (на поточном шифре), если:

- по побочным каналам утечки получена информация о возможных вариантах символов сообщения
- мощность ключевого множества меньше мощности алфавита открытого текста
- знаки ключевой последовательности не распределены равновероятно
- многократно используется один и тот же ключ шифрования

Описание алгоритма

- Составление словарей s-грамм
- Отбор «ограниченных» отрезков сообщения длиной s символов
- Построение вариантов восстановления отрезка
- Отбор осмысленных вариантов восстановления по словарю
- Восстановление отрезков сообщения

Описание алгоритма

- Критерий отбора: *среднее геометрическое* значение отрезка:

$$L_i = \sqrt[s]{l_{i_1} \cdot \dots \cdot l_{i_s}} \le L_{\text{кр}}$$

(l_{i_j} – количество вариантов символа для *j*-ого знака в *i*-ой s-грамме)

- Максимально *число вариантов* восстановления для одного отрезка: $k = 2^{\beta \cdot s}$

ц	и	н	с	к	а	я	*	п	о	м	о	щ	ь	*
у	*	ж	г	р	щ	ъ	и	х	г	с	л	л	ч	ю
э	о	у	ь	я	м	й	т	в	л	ш	й	ю	й	б
ю	в	и	р	м	ъ	з	з	й	м	у	т	п	в	м
п	ж	о	ъ	х	т	ь	ж	ч	а	ъ	я	з	,	р
р	ш	к	,	з	ь	д	й	с	р	г	п	ц	е	у
й	,	б	ц	ч	в	ю	м	а	и	ж	к	р	н	о
х	щ	п	ж	э	у	и	ъ	э	,	м	с	к	п	з
г	ч	л	м	*	ф	т	е	м	з	в	ф	и	з	ъ
щ	м	.	у	в	к	а	щ	п	д	ы	,	т	я	п
в	э	а	с	б	д	г	ь	.	х	я	р	*	л	ш
а	а	э	ш	ы	ш	ч	к	щ	ч	к	в	й	о	ь
ф	т	з	в	ц	е	к	,	-	ж	*	щ	а	с	щ
я	п	ш	о	с	й	л	д	-	ъ	-	и	я	.	ы
ь	я	е	з	у	я	х	н	-	о	-	н	.	ж	.
-	б	ц	ы	о	.	б	у	-	у	-	ц	ш	-	-
-	ю	с	е	ь	з	-	г	-	ю	-	ю	-	-	-
-	д	ю	-	ф	-	-	-	-	в	-	д	-	-	-
-	х	г	-	к	-	-	-	-	ь	-	-	-	-	-
-	г	-	-	-	-	-	-	-	ф	-	-	-	-	-
-	ъ	-	-	-	-	-	-	-	т	-	-	-	-	-

Параметры алгоритма

- Длина отрезка текста *s*: **10-25** символов
- Граница среднего геометрического *L*: **8-16** символов
- Допустимая степень неоднозначности восстановления отрезка $k = 2^{0,1 \cdot s}$

Длина текста, s	10	15	20	25
Количество возможных вариантов, k	1-2	1-2	1-4	1-5

Параметры алгоритма

- Покрытие словарей

$$\text{покрытие} = 1 - \frac{n_s}{N_s}$$

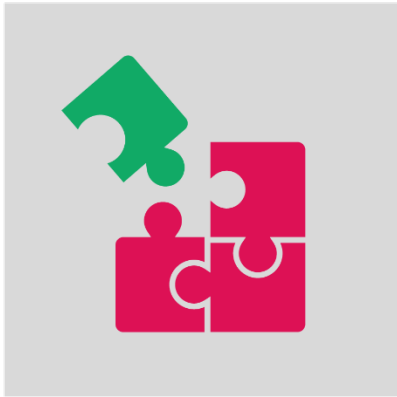
(N_s – исходный объём словаря s-грамм, n_s – число s-грамм, встречающихся один раз)

- Размер словаря с учетом покрытия

$$\check{N}_s = \frac{N_s}{1 - \frac{n_s}{N_s}}$$

- Энтропия s-грамм (бит на символ)

$$H_s = \log_2 \frac{\tilde{N}_s}{s}$$



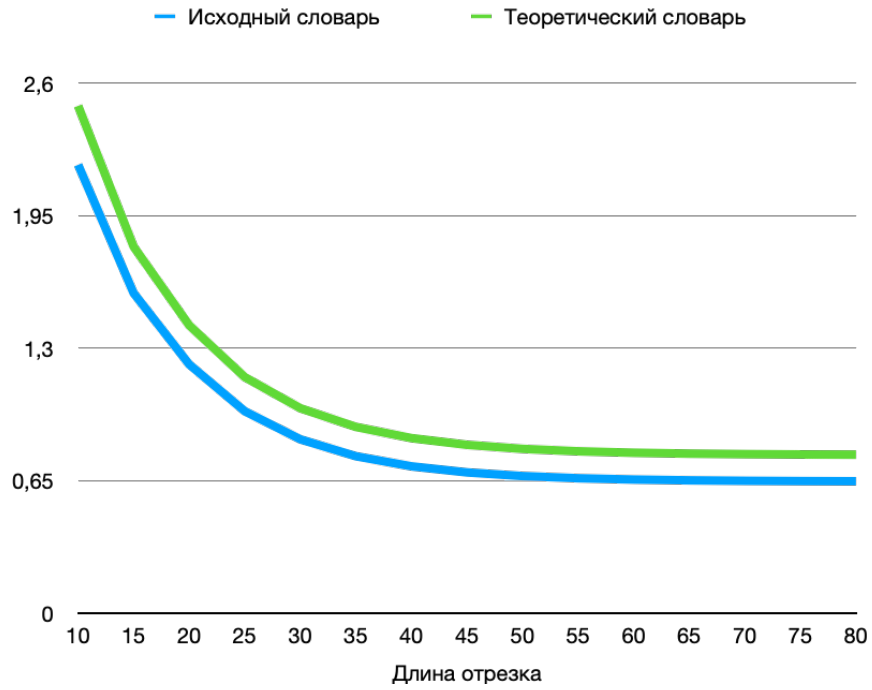
Энтропия s-грамм

- На текстовом материале русского языка объемом около 12 млн символов
- Учитываются только 32 буквы кириллического алфавита, знак пробела, точка и запятая

Длина отрезка	Объём словаря	Энтропия (эксп.) бит/симв	Покрытие (%)	Теоретический объём словаря	Энтропия (теор.) бит/симв
10	6,2 млн	2,26	19,95	~31 млн	2,49
15	9,5 млн	1,57	7,21	~131 млн	1,80
20	10,4 млн	1,17	3,27	~317 млн	1,41
25	10,6 млн	0,93	2,07	~513 млн	1,16

Предельное значение энтропии

- Предполагается, что значения энтропии подчиняется линейному соотношению
- Лучшее приближение дает значение $k = 0{,}6$
- **Предельное значение энтропии** русских текстов на уровне $H = 0{,}78$ бит на символ



Математические свойства алгоритма

- Алфавит открытого текста фиксирован - *35 символов*
- Граница среднего для отрезка *L* фиксирована
- Количество вариантов знаков для всех отрезков распределено **независимо и случайно равновероятно** от 1 до 35

Вероятность появления ограниченных отрезков

Определим характеристики *i*-го отрезка сообщения:

- $S_i = \sum_{j=1}^{s} \log_L l_{i_j}$,
- $\mu = \frac{1}{35}\sum_{k=1}^{35} \log_L k$
- $\sigma^2 = \frac{1}{35}\sum_{k=1}^{35} \log^2{}_L\, k - \left(\frac{1}{35}\sum_{k=1}^{35} \log_L k\right)^2$
- Φ – функция стандартного нормального распределения

для любого *i*

- ***Утверждение 1:***

Пусть длина отрезка сообщения $s \to \infty$. Тогда вероятность, что среднее геометрическое *i*-ого отрезка не превосходит заданной границы L:

$$\lim_{s\to\infty} P\left(\sqrt[s]{l_{i_1} \cdot l_{i_2} \cdot \dots \cdot l_{i_s}} \le L\right) = \Phi\left(\frac{s - s\mu}{\sqrt{s}\sigma}\right)$$

- ***Утверждение 2:***

Ожидаемое количество ограниченных отрезков длины s в сообщении длиной N символов:

$$(N - s + 1) \cdot P\left(\sqrt[s]{l_{i_1} \cdot l_{i_2} \cdot \dots \cdot l_{i_s}} \le L\right)$$

Вероятность появления ограниченных отрезков

Утверждение 3:

- *Условная вероятность* появления ограниченного отрезка при условии, что предыдущий отрезок ограничен

$$\lim_{s\to\infty} P(S_i < s \mid S_{i-1} < s) = \frac{\iint_0^s f(x,y)dxdy}{\Phi(\frac{s-s\mu}{\sqrt{s}\sigma})}$$

- *Ожидаемая средняя геометрическая величина* следующего отрезка при условии ограниченности предыдущего

$$\lim_{s\to\infty} E(S_i \mid S_{i-1} < s) = \mu - \frac{s-1}{s}\sigma\frac{\varphi\left(\frac{s-s\mu}{\sqrt{s}\sigma}\right)}{\Phi\left(\frac{s-s\mu}{\sqrt{s}\sigma}\right)}$$

- $f(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu)^2}{\sigma^2} - \rho\frac{2(x-\mu)(y-\mu)}{\sigma^2} + \frac{(y-\mu)^2}{\sigma^2}\right)}$
- **коэффициент корреляции** S_i и S_{i-1}

$$\rho = \frac{s-1}{s}$$

Вероятность появления ограниченных отрезков

Утверждение 4:

- Вероятность того, что все отрезки сообщения одновременно являются ограниченными, стремится к многомерному нормальному распределению

$$\lim_{s\to\infty} P(S_1 < s, S_2 < s, \dots, S_{N-s+1} < s) = \lim_{s\to\infty} P(\vec{S} < s) = N(\vec{\theta}, \Sigma s)$$

- Вектор средних значений

$$\vec{\theta} = \begin{pmatrix} s\mu \\ s\mu \\ \dots \\ s\mu \end{pmatrix}$$

- Ковариационная матрица

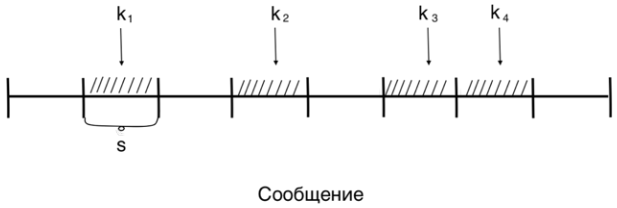
$$\Sigma s = \begin{bmatrix} s\sigma^2 & \cdots & \sigma_{S_1, S_{N-s+1}} \\ \vdots & \ddots & \vdots \\ \sigma_{S_{N-s+1}, S_1} & \cdots & s\sigma^2 \end{bmatrix}$$

- на пересечении k-ой строчки и z-го столбца - значения ковариации отрезков

$$\sigma_{S_k, S_z} = \operatorname{cov}(S_k, S_z)$$
$$= \begin{cases} (s - z + k)\sigma^2, \text{если } z - k < s \\ 0, \quad \text{иначе.} \end{cases}$$

Распределение осмысленных текстов

- $N = m^s$ – общее количество s-грамм, которые могут быть построены из символов алфавита мощностью *m=35* символов
- D – число осмысленных текстов длины *s* в том же алфавите (эта величина оценивается как 2^{Hs}, где H – энтропия s-грамм)
- $n_i = l_i^s$ – число возможных вариантов восстановления для *i*-го отрезка сообщения
- k_i – количество *осмысленных* вариантов восстановления для *i*-го восстанавливаемого отрезка сообщения среди n_i, где $l_i = \sqrt[s]{l_{i_1} \cdot \cdots \cdot l_{i_s}}$ – среднее число вариантов знаков для неизвестного символа *i*-го участка сообщения



- При этом предполагается, что **истинный вариант восстановления всегда присутствует** в выборке, то есть $n_i - 1$ – это ложные варианты

Распределение осмысленных текстов

- Вероятность того, что в выборке из n_i различных вариантов восстановления, ровно k_i вариантов окажутся осмысленными, описывается *гипергеометрическим распределением*

$$P(k_i) = \frac{C_D^{k_i} C_{N-D}^{n_i - k_i}}{C_N^{n_i}}$$

Утверждение 5:

- Если восстановлению подвергается участок сообщения на русском языке (мощность алфавита – 35 символов) длиной s символов со среднем числом вариантов знаков l_i, то вероятность появления k_i осмысленных вариантов восстановления данного отрезка:

$$P(k_i) = \frac{2^{H_s \cdot s}! \cdot (35^s - 2^{H_s \cdot s})! \cdot l_i^s! \cdot (35^s - l_i^s)!}{k_i! \cdot (2^{H_s \cdot s} - k_i)! \cdot (l_i^s - k_i)! \cdot (35^s - 2^{H_s \cdot s} - l_i^s + k_i)! \cdot 35^s!}$$

Наиболее **вероятное число осмысленных текстов,** которое будет найдено при восстановлении

	10	**15**	**20**	**25**
8	3	1	1	1
10	24	1	1	1
12	143	2	1	1
14	667	11	1	1
16	2534	80	1	1

Оптимальный параметр алгоритма - **16 символов**

Распределение осмысленных текстов

Предельное распределение числа осмысленных текстов возникает, когда $s \to \infty$. Тогда параметры гипергеометрического распределения $N = 35^s \to \infty, D = 2^{H \cdot s} \to \infty, n = l^s \to \infty, l \leq 35$. Вид предельного распределения зависит от количества осмысленных текстов в выборке.

- ***Утверждение 6:***

Вероятность найти ровно 1 осмысленный текст (истинный) при $s \to \infty$:

$$P(1) \approx e^{-2\left(\frac{l \cdot 2^H}{35}\right)^s}$$

- ***Утверждение 7:***

Если количество осмысленных текстов $k = 2^{s \cdot \beta}$, $l \leq 21, H \approx 0.8, \beta = 0.1$, тогда вероятность получить *k* осмысленных текстов при восстановлении отрезка длиной $s \to \infty$:

$$P\left(k = 2^{\beta s}\right) \approx \frac{1}{\sqrt{\pi}} 2^{\left(\left(H-\beta+\log_2 \frac{l}{35}\right)s+\log_2 e\right)\cdot 2^{\beta s}-\frac{\beta}{2}s-\frac{1}{2}}$$

Вероятность успеха

L	10	15	16*	20	25
<8	0,014	0,006	0,005	0,002	0,001
<10	0,016	0,065	0,060	0,041	0,027
<12	0,016	0,213	0,243	0,219	0,193
<14	0,016	0,256	0,512	0,514	0,515
<16	0,016	0,256	0,745	0,769	0,795
<18	0,016	0,256	0,887	0,912	0,935
<20	0,016	0,256	0,956	0,972	0,984
<22	0,016	0,256	0,984	0,992	0,996
<24	0,016	0,256	0,995	0,998	0,999

Оценка сложности реализации

- Этап составления словарей реализуется заранее и не входит в общую оценку трудоемкости
- Количество вариантов восстановления *для одного отрезка*, проверяемых на присутствие в словаре, *в среднем* $n = l^s$
- Бинарный поиск по словарю $O(log_2 D)$
- Поскольку трудоемкость относительно невелика, эффективность определяется *средней долей восстановленной информации*

Спасибо!