2.4. Криптоанализ шифра Виженера

В этом разделе будут описаны некоторые методы криптоанализа шифра Виженéра.

В случае рассматриваемого шифра, прежде всего, нужно найти длину ключевого слова, которую мы обозначим через l. Приведем два метода, которые могут быть использованы для этого. Первый из них — это *тест Казиски,* а второй — метод *индексов совпадения.*

Тест Казиски был предложен Ф. Казиски в 1863 году. Он основан на следующих соображениях: если два одинаковых отрезка открытого текста получаются один из другого сдвигом на величину, кратную длине ключевого слова, то они при шифровании перейдут в одинаковые отрезки шифртекста. Эти аргументы используются следующим образом: если в шифртексте имеются два фрагмента длины три или больше, то весьма вероятно, что они соответствуют одинаковым отрезкам открытого текста.

Тест Казиски применяется так. Ищем в шифртексте пары одинаковых отрезков длины три или больше. Находим расстояние между стартовыми позициями этих отрезков. Если мы найдем несколько таких расстояний d_1, d_2,… , то можно предположить, что искомое число l является делителем каждого из этих чисел и, следовательно, делителем их наибольшего общего делителя.

Дальнейшие аргументы для нахождения числа l могут быть получены с помощью *индексов совпадения*. Это понятие было введено В. Фридманом в 1920 году.

ОПРЕДЕЛЕНИЕ. *Пусть* $\mathbf{x}$ *— строка, составленная из букв некоторого алфавита. Индексом совпадения строки* $\mathbf{x}$ *называется вероятность того,*

что две случайно выбранных из этой строки буквы являются одинаковыми.

Индекс совпадения строки $\mathbf{x}$ будем обозначать через $I(\mathbf{x})$.

Будем предполагать, что рассматриваемый алфавит содержит m букв, пронумерованных числами $0, 1, \ldots, m-1$. Рассмотрим строку $\mathbf{x}$, содержащую n букв. Мы можем выбрать две буквы из этой строки,

$$C_n^2 = \frac{n(n-1)}{2}$$

способами. Предположим, что буква с номером i $(0 \le i \le m-1)$ встречается в этой строке f_i раз (напомним, что значение f_i называется частотой или, точнее, абсолютной частотой рассматриваемой буквы). Для произвольного $i(=0,1,\ldots,m-1)$ имеется $C_{f_i}^2$ способов выбрать одинаковые буквы с номером i. Отсюда получаем, что

$$I(\mathbf{x}) = \frac{\sum_{i=0}^{m-1} f_i(f_i-1)}{n(n-1)}.$$

Обозначим через p_0, $p_1,\ldots,p_{m-1}$ *вероятности* появления букв алфавита в этой строке, то есть *относительные частоты* f_0/n, f_1/n, …, f_{m-1}/n. Тогда при *достаточно большом значении* n, имеет место *приближенная* формула

$$I_c(\mathbf{x}) \approx \sum_{i=0}^{m-1} p_i^2,$$

так как вероятность дважды выбрать букву с номером 0 *приблизительно* равна p_0^2, букву с номером 1 — p_1^2, и так далее.

Предположим теперь, что строка составлена из букв естественного языка, например, русского или английского. Пользуясь таблицами, приведенными выше, получаем следующие значения индекса совпадения для

строки большой длины, являющейся : русский алфавит —**0.0553**, английский алфавит —**0.0644**. Те же значения будут получены и *для любого шифртекста,* полученного с помощью произвольного моноалфавитного шифра. Действительно, в этом случае вероятности отдельных букв поменяются местами, но *сумма* этих вероятностей останется неизменной.

Предположим, что дан шифртекст $\mathbf{y} = y_1 y_2 \ldots y_n$, полученный с помощью шифра Виженера. Сделаем предположение, что длина ключевого слова равна l. Введем в рассмотрение буквенный массив, состоящий из l строк и n/l столбцов. Запишем символы данного шифртекста по столбцам в этот массив. Строки этого массива обозначим через $\mathbf{y}_1$, $\mathbf{y}_2, \ldots, \mathbf{y}_l$. Если мы правильно нашли значение l, то каждая из строк $\mathbf{y}_1$, $\mathbf{y}_2, \ldots, \mathbf{y}_l$ получена из соответствующих подстрок исходного открытого текста с помощью шифра сдвига. Тогда значения $I(\mathbf{y}_i)$ $(1 \le i \le l)$ должны быть приближенно равны значению индекса совпадения для соответствующего языка (то есть **0.0553** для русского языка и **0.0644** для английского). Если мы неверно нашли значение l (точнее, если настоящее значение длины ключевого слова не является делителем числа l), то найденные подстроки уже будут иметь более случайную структуру, поскольку они получаются с помощью шифра сдвига с разными ключами. Заметим, что для полностью случайной строки в случае алфавита из m символов

$$I(\mathbf{x}) \approx m\left(\frac{1}{m}\right)^2 = \frac{1}{m}.$$

В случае русского языка это дает **0.03125**, в случае английского языка — **0.03856**. Для двух рассматриваемых языков это существенно отличается от приведенных выше значений и может позволить определить правильную длину ключевого слова (или подтвердить догадку, сделанную с помощью теста Казиски).

После определения длины ключевого слова нужно, разумеется, найти само это слово.

ОПРЕДЕЛЕНИЕ. *Пусть* $\mathbf{x}$, $\mathbf{y}$ *— строка, составленные из букв некоторого алфавита. Взаимным индексом совпадения этих строк называется вероятность того, что буква, случайно выбранная из первой строки, совпадает с буквой, случайно выбранной из второй строки.*

Взаимный индекс совпадения двух строк $\mathbf{x}$, $\mathbf{y}$ будем обозначать через $MI(\mathbf{x},\mathbf{y})$ ("*Mutual Index*"). Рассмотрим алфавит, состоящий из m букв. Пусть

$$\mathbf{x} = x_1 x_2 \ldots x_n, \quad \mathbf{y} = y_1 y_2 \ldots y_{n'},$$

p_0, $p_1, \ldots, p_{m-1}$ — вероятности того, что случайно выбранная из первой строки буква является соответственно первой, второй, …, m-й буквой алфавита, q_0, $q_1, \ldots, q_{m-1}$ — аналогичные характеристики для второй строки. Тогда, очевидно,

$$MI(\mathbf{x},\mathbf{y}) = \sum_{i=0}^{m-1} p_i q_i.$$

Обозначим через f_0, f_1, ,…, f_{m-1} частоты соответствующих букв алфавита в первой строке, а через f'_0, f'_1, ,…, f'_{m-1} — соответствующие характеристики для второй строки. Тогда

$$MI(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=0}^{m-1} f_i f'_i}{nn'}.$$

Возьмем произвольный шифр подстановки. Это означает, что выбрана произвольная подстановка $\varphi \in \mathfrak{S}_m$, рассматриваемая на множестве $\mathbb{Z}_m$, и буква алфавита с номером i заменяется буквой с номером $\varphi(i)$. Предположим, что строки $\mathbf{x}$ и $\mathbf{y}$ зашифрованы с помощью этого шифра. Пусть $\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$ — соответствующие строки шифртекста. Если, как и выше, p_0, $p_1, \ldots,$

p_{m-1} — вероятности выбора соответствующих букв алфавита из строки $\mathbf{x}$, q_0, $q_1,\ldots,q_{m-1}$ — аналогичные характеристики для строки $\mathbf{y}$. Для строки $\tilde{\mathbf{x}}$ эти вероятности равны $p_{\varphi^{-1}(0)}$, $p_{\varphi^{-1}(1)},\ldots,p_{\varphi^{-1}(m-1)}$. Тогда

$$MI(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \sum_{i=0}^{m-1} p_{\varphi^{-1}(i)} q_{\varphi^{-1}(i)} = \sum_{i=0}^{m-1} p_i q_i = MI(\mathbf{x},\mathbf{y}),$$

поскольку выписанные суммы отличаются только порядком следования слагаемых. Аналогично доказывается, что $I(\tilde{\mathbf{x}}) = I(\mathbf{x})$.

Мы получили важный факт: *индекс совпадения строки и взаимный индекс совпадения двух строк не меняются после шифровании этих строк произвольным шифром подстановки.*

Вернемся к анализу шифра Виженера. Пусть $K = (k_1,k_2,\ldots,k_l)$ — ключевое слово. Попробуем оценить величину $MI(\mathbf{y}_i,\mathbf{y}_j)$. Выберем случайным образом букву из строки $\mathbf{y}_i$ и букву из строки $\mathbf{y}_j$. Учтем, что каждая рассматриваемых строк шифруется с помощью шифра сдвига, первая — с ключом k_i, вторая — с ключом k_j. Вероятность того, что первая (вторая) из этих букв совпадает с первой буквой алфавита (нулем в $\mathbb{Z}_m$), равна p_{-k_i} (p_{-k_j}).Следовательно, вероятность того, что обе буквы совпадают с первой буквой алфавита, равна $p_{-k_i}p_{-k_j}$.

ЗАМЕЧАНИЕ. Здесь и далее в индексах операции выполняются в кольце $\mathbb{Z}_m$. Все равенства являются приближенными, поскольку вероятности появления букв в данных строках лишь приближенно равны среднестатистическим величинам.

Аналогично получаем: вероятность того, что обе буквы совпадают с второй буквой алфавита (элементом 1 в $\mathbb{Z}_m$), равна $p_{1-k_i}p_{1-k_j}$. Следовательно,

$$MI(\mathbf{y}_i, \mathbf{y}_j) \approx \sum_{h=0}^{m-1} p_{h-k_i} p_{h-k_j} = \sum_{h=0}^{m-1} p_h p_{r+k_i-k_j}.$$

Заметим, что правая часть этого выражения зависит только от *разности* $k_i - k_j$, которую будем называть *относительным сдвигом* строк $\mathbf{y}_i$ и $\mathbf{y}_j$.

Учтем следующую очевидную формулу:

$$\sum_{h=0}^{m-1} p_h p_{h+\alpha} = \sum_{h=0}^{m-1} p_h p_{h-\alpha}, \quad \alpha \in \mathbb{Z}_m,$$

то есть относительный сдвиг на величину α дает то же значение величины MI, что и сдвиг на величину $m - \alpha$. Это означает, что в случае алфавита из m букв, достаточно найти значения MI для значений $\alpha = 0, 1, \ldots, [m/2]$. Приведем эти значения для русского и английского языков.

Относительный сдвиг	Взаимный индекс
0	0.0553
1	0.0366
2	0.0345
3	0.0400
4	0.0340
5	0.0360
6	0.0326
7	0.0241
8	0.0287
9	0.0317
10	0.0265
11	0.0251
12	0.0244
13	0.0291

Относительный сдвиг	Взаимный индекс
14	0.0322
15	0.0244
16	0.0249

Таблица 10.

Взаимный индекс совпадения (русский язык)

Относительный сдвиг	Взаимный индекс
0	0.0644
1	0.0394
2	0.0319
3	0.0345
4	0.0436
5	0.0332
6	0.0363
7	0.0389
8	0.0338
9	0.0342
10	0.0378
11	0.0440
12	0.0387
13	0.0428

Таблица 11.

Взаимный индекс совпадения (английский язык)

Отметим следующее. Если относительный сдвиг в случае русского (английского) языка ненулевой, то значения взаимного индекса варьируются от **0.0241** до **0.0400** (от **0.0319** до **0.0440**), в то время как относи-

тельный индекс для нулевого относительного сдвига равен **0.0553** (соответственно **0.0644**). Это наблюдение позволяет делать правдоподобные предположения относительно разностей $\beta = K_i - K_j$.

Зафиксируем числа i, j, $1 \le i, j \le l$, $i \ne j$. Возьмем строки $\mathbf{y}_i$ и $\mathbf{y}_j$. Рассмотрим строки, получающиеся путем применения шифров сдвига со всеми возможными значениями ключа (то есть 0, 1, …, $m-1$) к строке $\mathbf{y}_j$. Обозначим получаемые строки через $\mathbf{y}_j^0$, $\mathbf{y}_j^1$, … , $\mathbf{y}_j^{m-1}$ ($m = 32$ в случае русского алфавита и $m = 26$ в случае английского алфавита). Теперь найдем соответствующие значения взаимного индекса совпадения $MI_c(\mathbf{y}_i, \mathbf{y}_j^\beta)$, $\beta = 0$, 1, … , $m-1$. При $\beta = \alpha$ взаимный индекс совпадения должен быть близким к наибольшему значению для данного языка, поскольку относительный сдвиг между $\mathbf{y}_i$ и $\mathbf{y}_j^\alpha$ равен нулю. При $\beta \ne \alpha$ рассматриваемая величина должна принимать существенно меньшие значения.

Пример. Рассмотрим шифртекст, полученный с использованием шифра Виженера:

```
СЮРСЕЫПЛУИДЮПЖДЭТВЙЛЧЯЭХИЪЩЭРОСШЕПКЭЭЕБТФЫКОО
ЮЮЩФЧЧЫХГАЬШУВОХХЯУВПЯУТСБРЫПВЧЫЯАНПЧОЕНЧХЩБП
КШФЖВПЫЭЙДХРШРЕАШХФЧЧЫХГОЬЮШХИЪЩЭРАГШЩЙПЮЮРЙЧ
ПФРБПЭЫЫПЗСГЕОЫДЮАСЫСШЯУЫДЕЬИЯШБШЙХВФОСАЧТЙМЕ
БРЩТЛПХЪГСБАУШАБМУДОСЯУСЕУЭУКХЭЧНЙНСТОМИСРЩТЯ
СЗУСНЭСБСОШЫЦТЕАТУСКППЪЙЛЪШЬООЫБЫЖЖЬЛГВЛФБАПК
ЮРГОУИШЧНАЬФОСИЬРЪЙСЧУОСАЫШТФХПЬЦМИЫЮЫБМЧСУОЗ
ЧЭЬНОУХШПЛЭЭЬНСВЪЫПМЧУЬМОАХСПКПЪШПМПЭТОАОВЮФБ
ПАОИНЭБЦМСОЯЬГСФЬБЗИЪЩЗФ
```

Прежде всего, воспользуемся тестом Казиски. Триграмма **"ИЪШ"** появляется в тексте три раза, начинаясь с букв с номерами **25**, **120** и **380**

(подчеркнуто в шифртексте). Разности между этими числами равны **95** и **260**. Наибольший общий делитель двух последних чисел равен пяти. Поэтому правдоподобной является гипотеза, что $l = 5$.

Посмотрим, дает ли подсчет индексов совпадения такой же вывод. При $l = 1$ индекс совпадения (всей строки шифртекста) равен **0.0360**.

При $l = 2$ текст разбивается на две строки

СРЕПУДПДТЙЧЭИШРСЕКЭБФКОЮФЧХАШВХЯВЯТБЫВЫАПОНХБ
КФВЫЙХШЕШФЧХОЮХЪЭАШЙЮРЧФБЭЫЗГОДАЫШУДЬЯБЙВОАТМ
БЩЛХГБУАМДСУЕЭКЭННТМСЩЯЗСЭБОЫТАУКПЙЪЬОБЖЬГЛБП
ЮГУШНЬОИРЙЧОАШФПЦИЮБЧУЗЭНУШЛЭНВЫМУМАСКЪППТАВФ
ПОНБМОЬСЬЗЪЗ

и

ЮСЫЛИЮЖЭВЛЯХЪЭОШПЭЕТЫОЮЩЧЫГЬУОХУПУСРПЧЯНЧЕЧЩП
ШЖПЭДРРАХЧЫГЬШИШРГЩПЮЙПРПЫПСЕЫЮССЯЫЕИШШХФСЧЙЕ
РТПЪСАШБУОЯСУУХЧЙСОИРТСУНССШЦЕТСПЪЛШОЫЬЖЛВФАК
РОИЧАФСЬЪСУСЫТХЬМЫЫМСОЧЬОХПЭЬСЪПЧЬОХППШМЭООЮБ
АИЭЦСЯГФБИШФ

с индексами совпадения соответственно **0.0323** и **0.0417**. При $l = 3$ индексы совпадения равны соответственно **0.0389**, **0.0337** и **0.0359**, при $l = 4$ — **0.0309**, **0.0425**, **0.0340** и **0.0373**, при $l = 5$ —**0.0635**, **0.0561**, **0.0435**, **0.0547** и **0.0526**, при $l = 6$ — **0.0332**, **0.0382**, **0.0283**, **0.0397**, **0.0322** и **0.0486**. Заметное увеличение индексов совпадения при $l = 5$ также говорит в пользу нашей гипотезы.

Теперь попробуем найти величины относительных сдвигов. Приведем таблицу 320 значений величин $MI_c(\mathbf{y}_i, \mathbf{y}_j^{\beta})$, $1 \le i < j \le 5$, $0 \le \beta \le 31$. В ней значения индекса указаны в порядке возрастания величины β от нуля до 31. В этих значениях отброшены начальные символы **0.0**. Для каждой

пары (i, j) выберем наибольшее значение индекса. Это значение подчеркнуто.

i	j	Значение $MI_c(\mathbf{y}_i, \mathbf{y}_j^{\beta})$
1	**2**	**391 306 405 398 313 402 313 279 294 277 232 216 296 220 218 178 289 279 284 241 305 184 317 362** 211 414 346 255 455 395 294 630
1	**3**	322 495 358 339 436 362 268 319 298 348 325 211 263 232 300 255 220 189 305 260 284 268 312 293 294 280 272 355 462 267 419 390
1	**4**	383 260 312 350 206 298 429 263 435 409 315 438 431 362 563 403 338 341 301 293 338 258 213 220 185 185 216 263 220 239 289 246
1	**5**	336 289 325 306 284 255 362 331 211 447 372 270 391 390 267 604 370 320 367 376 293 324 313 303 296 216 263 173 336 232 248 130
2	**3**	405 351 525 374 343 499 332 301 353 249 341 305 225 268 237 237 244 192 189 268 220 249 248 261 310 317 272 312 398 400 345 429
2	**4**	239 267 242 305 334 229 355 438 289 469 443 374 532 412 417 587 370 336 421 249 268 319 216 234 235 161 189 197 204 209 227 232
2	**5**	190 263 235 294 272 234 308 381 268 291 474 322 348 464 355 325 616 364 332 466 319 306 345 249 303 296 211 249 189 272 230 227
3	**4**	255 277 360 220 370 376 300 424 358 376 506 424 424 459 357 369 391 306 348 306 275 218 251 227 204 223 229 204 237 229 204 293

i	j	Значение $MI_c(\mathbf{y}_i, \mathbf{y}_j^{\beta})$
3	**5**	**268 272 291 244 284 362 332 293 353 319 351 443** 343 339 <u>547</u> 348 358 376 315 407 353 280 251 298 301 261 215 256 244 282 206 206
4	**5**	384 <u>521</u> 409 452 485 353 388 433 277 383 327 201 303 251 249 265 206 239 251 222 258 235 209 222 296 235 265 360 310 282 400 329

Таблица 12.

Значения взаимных индексов совпадения

Выпишем уравнения (в кольце $\mathbb{Z}_{32}$), связывающие неизвестные значения K_1, K_2, K_3, K_4, K_5, отвечающие наибольшим из шести подчеркнутых значений:

$$K_1 - K_2 = 31, \quad K_2 - K_5 = 16, \quad K_1 - K_5 = 15,$$
$$K_2 - K_4 = 15, \quad K_1 - K_4 = 14, \quad K_3 - K_5 = 14.'$$

Эти уравнения позволяют выразить все значения K_i через K_1:

$$K_2 = K_1 + 1, \quad K_3 = K_1 + 31, \quad K_4 = K_1 + 18, \quad K_5 = K_1 + 17.$$

Таким образом, предполагается, что ключ имеет следующий вид

$$(K_1, K_1 + 1, K_1 + 31, K_1 + 18, K_1 + 17)$$

для некоторого значения $K_1 \in \mathbb{Z}_{32}$. Теперь число вариантов для ключа равно 32, в отличие от первоначальных $32^5 = 33554432$. Перебирая возможные значения K_1, находим то, при котором получается осмысленный текст: $K_1 = 15$, ключ имеет в терминах $\mathbb{Z}_{32}$ вид $(15,16,14,1,0)$, или слово **"ПРОБА"** в терминах алфавита. Окончательный вид открытого текста следующий:

Во время этих обедов Филипп Филиппович окончательно получил звание божества. Пес

становился на задние лапы и жевал пиджак, пес изучил звонок Филиппа Филипповича - два полнозвучных отрывистых хозяйских удара, и вылетал с лаем встречать его в передней. Хозяин вваливался в чернобурой лисе, сверкая миллионом снежных блесток, пахнущий мандаринами, сигарами, духами, лимонами, бензином, одеколоном, сукном, и голос его, как командная труба, разносился по всему жилищу.[1]

В заключение заметим, что не все уравнения, получаемые указанным способом, являются верными. Например, одно из уравнений имеет вид

$$K_3 - K_4 = 10.$$

В действительности уравнение должно иметь вид

$$K_3 - K_4 = 13.$$

Оно определяется вторым по величине значением индекса, подчеркнутым в таблице волнистой чертой.

[1] М.А.Булгаков. *Собачье сердце.*