http://rosoperator.ru/rameev100/
Всероссийская научно-техническая конференция, посвященная 100-летию со дня рождения одного из основоположников советской вычислительной техники Б.И. Рамеева, на тему: «Информационно-управляющие и телекоммуникационные системы специального назначения». Секция «Биометрическая поддержка криптовалют, блокчейн реестров, облачных сервисов», проводится в конференц-зале Технопарка высоких технологий «Рамеев», доклад 16 мая 2018 года с 12^{30} до 12^{45} (ул. Центральная, д. 1), г. Пенза.

Карпов А.П., Юнин А.П.

Условия корректного вычисление энтропии осмысленных длинных паролей в пространстве сверток Хэмминга с эталонными текстами на русском и английском языках

Аннотация.

Актуальность и цели. Целью работы является повышение корректности вычисления энтропии длинных кодов с зависимыми разрядами, являющимися осмысленными легко запоминаемыми паролями на родном языке пользователя.

Материалы и методы. Классические процедуры Шеннона не могут быть использованы, так как требуют использования огромного статистического материала. Для сокращения затрат вычислительных ресурсов используется отображение кодов в нормированное пространство сверток Хэмминга.

Результаты. Показано, что результаты вычислений являются более корректными, если отказаться от побитного сложения по модулю два при вычислении сверток Хэмминга. Предложено использовать свертывание данных по модулю 8, так как кодирование паролей и эталонных текстов выполняются в 8-ми битной кодировке. Более того, корректное преобразование данных может быть выполнено только при использовании кода длинного пароля, свертываемого с эталонным текстом на родном языке пользователя.

Выводы. В пространстве сверток Хэмминга легко вычислим прирост стойкости длинных легко запоминаемых паролей со смыслом к атакам подбора, возникающего из-за периодической смены регистра ввода длинного пароля.

Ключевые слова: энтропия длинных кодов с зависимыми разрядами, регуляризация вычислений, многообразие сверток Хэмминга, требования к перекодировке данных перед их свертыванием по Хэммингу.

Проблема вычисления энтропии длинных кодов с зависимыми разрядами

Если пытаться вычислять энтропию длинных кодов по Шеннону, то мы сталкиваемся с задачей экспоненциальной вычислительной сложности. Так для кодов длинной 256 бит, полученных от программного генератора псевдослучайных чисел возникает 2^{256} состояний. Произведение «Война и мир» в 4 томах Льва Толстого имеет1640 страниц, 2000 знаков на странице дает 2^{22} знаков. Пользуясь как эталонным текстом русского языка произведением «Война и мир» по Шеннону мы можем оценивать пароли длинной до 176 бит или 22 знака. Для оценки пароля длинной в 32 случайных знака потребуется 2^{130} произведений на русском языке размерами сопоставимыми с 4-мя томами «Войны и мира». Все оцифрованные русскоязычные источники на содержат такой объем информации. Даже если бы такой эталон русскоязычного текста существовал, его анализ на обычном современном компьютере может занять тысячи лет машинного времени.

Проблема состоит в том, что руководствуясь Шенноном приходится обрабатывать большие массивы данных и ждать появления редких событий. Положение меняется, если мы из пространства обычных кодов переходим в пространство расстояний Хэмминга [1, 2,

3]. Для кодов длинной 256 бит расстояний Хэмминга меняется в интервале 0≤ h ≤256, итого 257 состояний:

$$h = 256 - \sum_{i=1}^{256} ("c_i") \oplus ("x_i") \qquad (1),$$

где "c_i" - разряд кода длинного пароля, "x_i" - этот же разряд кода эталонного текста.

В работах [4, 5, 6] показано, что свертка Хэмминга может быть выполнена не только по модулю два. Для того, что бы обобщить результаты сверток и сделать их сопоставимыми, нормируем интервал, в котором могут меняться расстояния Хэмминга:

$$\tilde{h} = \frac{h}{\max(h)} \qquad (2).$$

В этом случае нормированные расстояния всех сверток Хэмминга всегда будут находиться в интервале от 0 до 1. Для примера на рисунке 1 даны распределения нормированных расстояний Хэмминга для эталонных текстов на русском и английском языках.

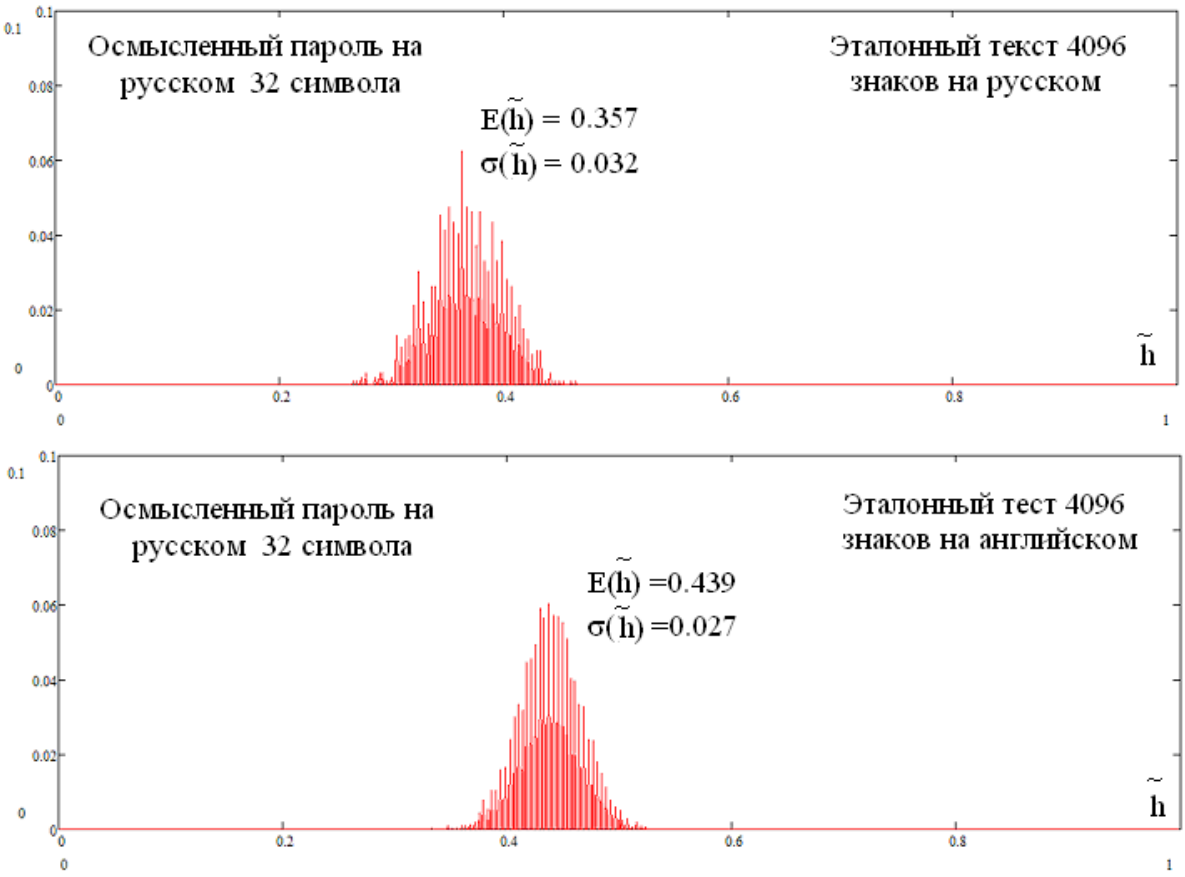

Рис. 1. Распределения расстояний Хэмминга при свертывании кода длинного осмысленного пароля с эталонными текстами на русском и английском языках

Из рисунка 1 видно, что распределение расстояний Хэмминга при тестирование пароля на русском языке ближе к состоянию $\tilde{h} = 0$, то есть, подбирая пароль сочетаниями фраз на русском мы получим меньшую вероятность ошибок второго рода. В итоге оценка энтропии пароля в среде MathCAD дает величину:

$$-\log\left(\text{pnorm}\left(\frac{1}{256}, 0.357, 0.032\right), 2\right) = 92.628 \text{ бит}$$

Если мы будем пытаться осуществить атаку, подбирая пароль на русском английскими фразами, то получим очень большую оценку энтропии:

$$-\log\left(\text{pnorm}\left(\frac{1}{256},0.439,0.027\right),2\right) = 192.661 \text{ бит}$$

Смысл подобных оценок понятен, пароль на русском языке следует подбирать, пользуясь фрагментами текстов на русском языке.

Следует отметить, что приведенные выше оценки являются слишком оптимистичными. Это обусловлено тем, что при вычислениях мы не принимали в расчет 8-битную кодировку символов. Учет 8-ми битной структуры кодов ASCII приводит к необходимости вычислять свертки Хэмминга по модулю восемь:

$$h_8 = 256 \cdot 32 - \sum_{i=1}^{32} \left("c_i, c_{i+1}, ..., c_{i+8}"\right) \oplus_8 \left("x_i, x_{i+1}, ..., x_{i+8}"\right) \qquad (3).$$

В итоге мы получаем более реалистичные распределения расстояний Хэмминга, приведенные на рисунке 2.

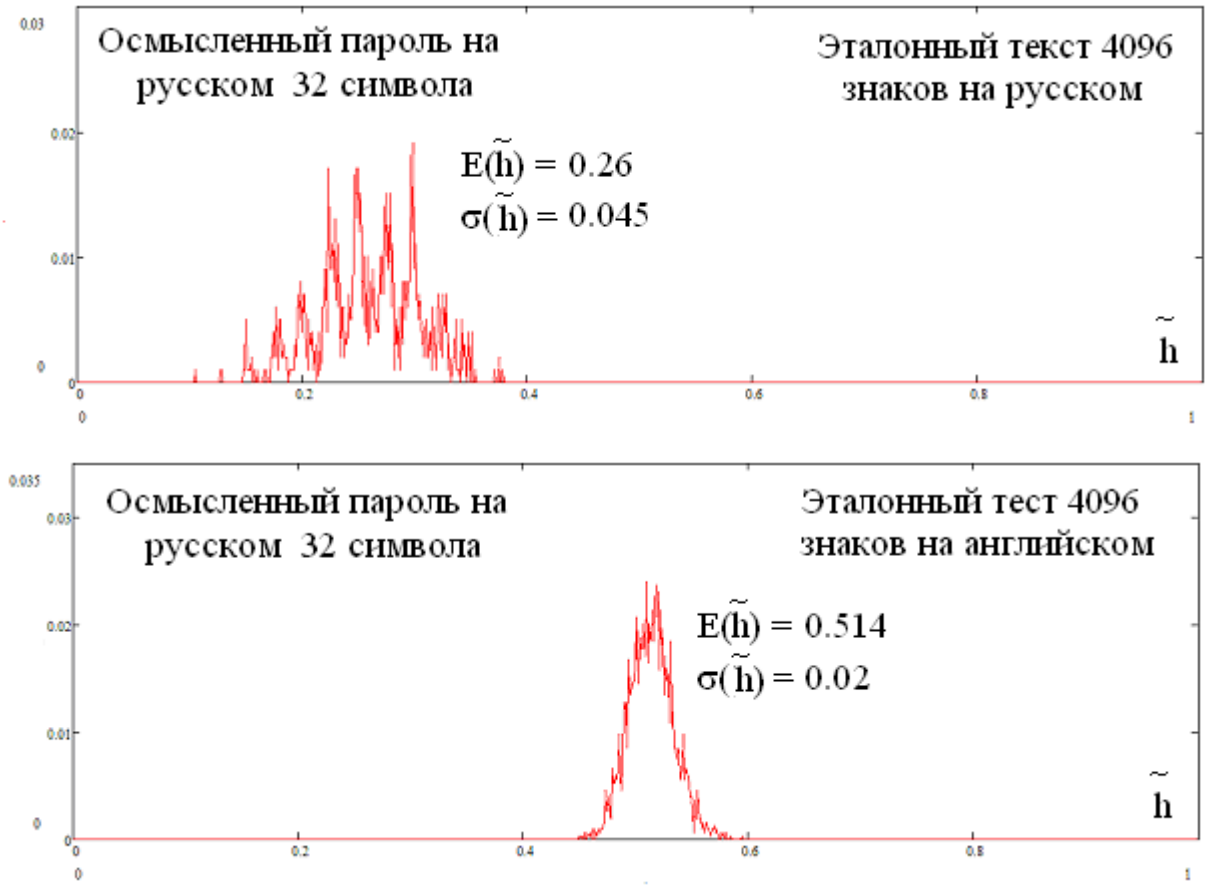

Рис. 2 Распределение расстояний Хэмминга в 8-ми битной системе счисления со свертыванием данных по модулю 8

Оценка энтропии для осмысленного пароля на русском при его тестировании тоже на русском получается ниже.

$$-\log\left(\text{pnorm}\left(\frac{1}{256 \cdot 32},0.26,0.045\right),2\right) = 27.954 \text{ бит}$$

Если мы такую же оценку выполняем применяя сочетания слов на английском, то получаем увеличение энтропии:

$$-\log\left(\text{pnorm}\left(\frac{1}{256\cdot 32},0.514,0.02\right),2\right) = 482.228 \text{ бит}$$

И в том и в другом случае получаются гораздо более реалистичные оценки энтропии. И в двоичной и восьмиричной системах сверток Хэмминга мы наблюдаем дефект вычислений (неустойчивость метода) когда тестируем пароль на другом языке. В восьмиричной системе счисления этот дефект усилился.

Наличие этого дефекта связано с тем, что ASCII кодировки имеют компактное расположение кодов букв латиницы и кодов букв кириллицы. Оба этих алфавита имеют расстояние между центрами групп «латиницы» и «кириллицы» 224-96=128 (7 бит). Именно это обстоятельство и приводит к расхождению математических ожиданий расстояний Хэмминга распределений рисунка 1 и рисунка 2. Эту ситуацию иллюстрирует рисунок 3 и рисунок 4.

Первая половина таблицы кодов ASCII

символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код
	32	00100000	8	56	00111000	P	80	01010000	h	104	01101000
!	33	00100001	9	57	00111001	Q	81	01010001	i	105	01101001
"	34	00100010	:	58	00111010	R	82	01010010	j	106	01101010
#	35	00100011	;	59	00111011	S	83	01010011	k	107	01101011
$	36	00100100	<	60	00111100	T	84	01010100	l	108	01101100
%	37	00100101	=	61	00111101	U	85	01010101	m	109	01101101
&	38	00100110	>	62	00111110	V	86	01010110	n	110	01101110
'	39	00100111	?	63	00111111	W	87	01010111	o	111	01101111
(	40	00101000	@	64	01000000	X	88	01011000	p	112	01110000
)	41	00101001	A	65	01000001	Y	89	01011001	q	113	01110001
*	42	00101010	B	66	01000010	Z	90	01011010	r	114	01110010
+	43	00101011	C	67	01000011	[	91	01011011	s	115	01110011
,	44	00101100	D	68	01000100	\	92	01011100	t	116	01110100
-	45	00101101	E	69	01000101	]	93	01011101	u	117	01110101
.	46	00101110	F	70	01000110	^	94	01011110	v	118	01110110
/	47	00101111	G	71	01000111	_	95	01011111	w	119	01110111
0	48	00110000	H	72	01001000	`	96	01100000	x	120	01111000
1	49	00110001	I	73	01001001	a	97	01100001	y	121	01111001
2	50	00110010	J	74	01001010	b	98	01100010	z	122	01111010
3	51	00110011	K	75	01001011	c	99	01100011	{	123	01111011
4	52	00110100	L	76	01001100	d	100	01100100	|	124	01111100
5	53	00110101	M	77	01001101	e	101	01100101	}	125	01111101
6	54	00110110	N	78	01001110	f	102	01100110	~	126	01111110
7	55	00110111	O	79	01001111	g	103	01100111	□	127	01111111

Рис. 3. первая половина вариантов восьми битной ASCII кодировки символов (заливкой отмечены группы символов, используемых при кодировке текстов на английском)

На величину стандартного отклонения распределения расстояний Хэмминга прежде всего влияет компактность кодировки групп символов (отсутствие разрывов между кодами). Как следствие, сделать процедуры вычисления сверток Хэмминга более устойчивыми удается перекодировками, которые ликвидируют пробелы между кодами в группах «латиница» для текстов на английском и «кириллица» для текстов на русском. Часто используемые в текстах знаки препинания должны иметь коды в группе символов в соотвествии с вероятностью их появления в тексте. Группировка кодов и их упорядочивание по частоте появления символов являются мощными методами структурной регуляризации вычислений энтропии.

Вторая половина таблицы кодов ASCII

символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код
Ђ	128	10000000		160	10100000	А	192	11000000	а	224	11100000
Ѓ	129	10000001	Ў	161	10100001	Б	193	11000001	б	225	11100001
‚	130	10000010	ў	162	10100010	В	194	11000010	в	226	11100010
ѓ	131	10000011	Ј	163	10100011	Г	195	11000011	г	227	11100011
„	132	10000100	¤	164	10100100	Д	196	11000100	д	228	11100100
…	133	10000101	Ґ	165	10100101	Е	197	11000101	е	229	11100101
†	134	10000110	¦	166	10100110	Ж	198	11000110	ж	230	11100110
‡	135	10000111	§	167	10100111	З	199	11000111	з	231	11100111
€	136	10001000	Ё	168	10101000	И	200	11001000	и	232	11101000
‰	137	10001001	©	169	10101001	Й	201	11001001	й	233	11101001
Љ	138	10001010	Є	170	10101010	К	202	11001010	к	234	11101010
‹	139	10001011	«	171	10101011	Л	203	11001011	л	235	11101011
Њ	140	10001100	¬	172	10101100	М	204	11001100	м	236	11101100
Ќ	141	10001101		173	10101101	Н	205	11001101	н	237	11101101
Ћ	142	10001110	®	174	10101110	О	206	11001110	о	238	11101110
Џ	143	10001111	Ї	175	10101111	П	207	11001111	п	239	11101111
ђ	144	10010000	°	176	10110000	Р	208	11010000	р	240	11110000
‘	145	10010001	±	177	10110001	С	209	11010001	с	241	11110001
’	146	10010010	І	178	10110010	Т	210	11010010	т	242	11110010
“	147	10010011	і	179	10110011	У	211	11010011	у	243	11110011
”	148	10010100	ґ	180	10110100	Ф	212	11010100	ф	244	11110100
•	149	10010101	µ	181	10110101	Х	213	11010101	х	245	11110101
–	150	10010110	¶	182	10110110	Ц	214	11010110	ц	246	11110110
—	151	10010111	·	183	10110111	Ч	215	11010111	ч	247	11110111
□	152	10011000	ё	184	10111000	Ш	216	11011000	ш	248	11111000
™	153	10011001	№	185	10111001	Щ	217	11011001	щ	249	11111001
љ	154	10011010	є	186	10111010	Ъ	218	11011010	ъ	250	11111010
›	155	10011011	»	187	10111011	Ы	219	11011011	ы	251	11111011
њ	156	10011100	ј	188	10111100	Ь	220	11011100	ь	252	11111100
ќ	157	10011101	Ѕ	189	10111101	Э	221	11011101	э	253	11111101
ћ	158	10011110	ѕ	190	10111110	Ю	222	11011110	ю	254	11111110
џ	159	10011111	ї	191	10111111	Я	223	11011111	я	255	11111111

Рис. 4. вторая половина восьми битной ASCII кодировки символов (заливкой отмечены группы символов, используемых при кодировке текстов на русском)

Таким образом, оценку энтропии в пространстве сверток Хэмминга можно сделать еще более устойчивой, если осуществлять предварительную перекодировку символов ASCII кодировки по специальную кодировку, обеспечивающую минимизацию значения математического ожидания расстояний Хэмминга и их стандартного отклонения.

ЛИТЕРАТУРА:

1. Иванов А.И., Ефимов О.В., Фунтиков В.А. Оценка усиления стойкости коротких цифровых паролей (PIN кодов) при их рукописном воспроизведении / «Защита информации. INSIDE» № 1, 2006 г., с. 55-57.
2. **.** Малыгин А.Ю., Волчихин В.И., Иванов А.И., Фунтиков В.А. Быстрые алгоритмы тестирования нейросетевых механизмов биометрико-криптографической защиты информации /Пенза-2006 г., Издательство Пензенского государственного университета., 161 с.
3. ГОСТ Р 52633.3-2011 «Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора».
4. Юнин А.П., Корнеев О.В. Оценка энтропии легко запоминаемых, длинных паролей со смыслом в ASCII кодировке для русского и английского языков

//Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора». Труды научно-технической конференции кластера пензенских предприятий, обеспечивающих БЕЗОПАСНОСТЬ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ. Том 10, Пенза-2016, с. 40-42 (http://пниэи.рф/activity/science/BIT/T10-p40.pdf)

5. Волчихин В.И., Иванов А.И., Юнин А.П., Малыгина Е.А. Многомерный портрет цифровых последовательностей идеального «белого шума» в свертках Хэмминга // «Известия высших учебных заведений. Поволжский регион. Технические науки.» 2017 г. №4. (стр.?? принята к опубликованию, в печати)
6. Иванов А.И. Многомерная нейросетевая обработка биометрических данных с программным воспроизведением эффектов квантовой суперпозиции. Издательство АО «ПНИЭИ», Пенза-2016 г., 133 с. Свободный доступ http://пниэи.рф/activity/science/BOOK16.pdf

Сведения об авторах:

Карпов Артем Павлович – специалист Пензенского филиала ФГУП «НТЦ «Атлас»

Юнин Алексей Петрович – специалист АО «ПНИЭИ»