ПРАКТИЧЕСКАЯ КРИПТОЛОГИЯ
ЛЕКЦИЯ 4
Специальность: 6.170101 – Бсіт
Лектор: Сушко С.А.

§5. ШИФР ВЕРНАМА

Пусть длина открытого текста, шифротекста и ключа одинакова, а их знаки выбираются из кольца вычетов Z_m. Уравнение шифрования

$$y_i \equiv (x_i + k_i) \bmod m,\ i = 1,2,...,n. \qquad (10)$$

переводит n-грамму открытого текста $(x_1, x_2, ..., x_n)$ на ключе $k = (k_1, k_2, ..., k_n)$ (гамме), в n-грамму шифрованного текста $(y_1, y_2, ..., y_n)$.

Существует немало модификаций этой схемы шифрования, наиболее известная из них – ***шифр Вернама*** за модулем m. Однократно используемый шифр Вернама по модулю 2 (суммирование по модулю 2 сводится к операции XOR) называют ***одноразовым шифровальным блокнотом***.

Теорема Шеннона о совершенной стойкости шифра Вернама. Шифр Вернама по модулю m – совершенно стойкий при случайном равновероятном выборе ключа из множества всех n-грам в алфавите Z_m.

Д о к а з а т е л ь с т в о. Мощность множества $\left|Z_{m,n}\right| = m^n$. Пусть все ключи $k = (k_1, k_2, ..., k_n)$ равновероятны и выбираются из множества $Z_{m,n}$ случайно. Вероятность выбрать ключ, который правильно переводит текст $x = (x_1, x_2, ..., x_n)$ в шифротекст $y = (y_1, y_2, ..., y_n)$, равна

$$p(k) = \frac{1}{m^n} = m^{-n}.$$

Тогда

$$p(y / x) = \sum_{\substack{k \in K \\ y = E_k(x)}} p(k) = m^{-n}.$$

По формуле Байеса получим

$$p(x / y) = \frac{p(x) \cdot p(y / x)}{\sum_{b \in Z_{m,n}} p(b) \cdot p(y / b)} =$$

$$= \frac{p(x) \cdot m^{-n}}{m^{-n} \sum\limits_{b \in Z_{m,n}} p(b)} = p(x).$$

По слухам "горячая линия» между США и СССР шифровалась с помощью одноразового блокнота.

Достоинства шифра Вернама	Особенности эксплуатации шифра Вернама
1. Все шифротексты равно-вероятны; 2. Абсолютная стойкость одноразового блокнота.	1. Каждый ключ можно использовать только один раз. 2. Гамма должна быть истинно случайной последовательностью чисел. 3. Длина ключа должна быть равна длине сообщения

На практике можно один раз физически передать носитель информации с длинным истинно случайным ключом, а потом по мере необходимости пересылать сообщения. На этом основана идея ***шифроблокнотов***: шифровальщик при личной встрече снабжается блокнотом, каждая страница которого содержит ключ. Такой же блокнот есть и у принимающей стороны. Использованные страницы уничтожаются.

Шифр, ключ которого по длине равен длине сообщения и может использоваться только один раз, неудобен при интенсивной эксплуатации.

§6. НЕКОТОРЫЕ СВЕДЕНИЯ ИЗ ТЕОРИИ ИНФОРМАЦИИ

В криптографию информация входит как величина, которую можно измерить (как измеряют длину в метрах, напряжение в вольтах). Такой подход дает теория информации, разработанная в 1948 р. К. Шенноном.

Информационная энтропия – это мера неопределённости или непредсказуемости информации, неопределённость появления какого-либо символа алфавита.

Например, в последовательности букв некоторого предложения разные буквы появляются с разной частотой, поэтому неопределённость появления для некоторых букв меньше, чем для других. Если же учесть, что некоторые буквенные n-граммы встречаются крайне редко, то неопределённость уменьшается еще сильнее.

Формальное определение энтропии. Пусть дискретная случайная величина X задана законом распределением вероятностей $X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$, где $x_1, x_2, \dots, x_n$ – значения дискретной величины, $\sum_{i=1}^{n} p_i = 1$. Тогда ***энтропия случайной величины*** X (***энтропия вероятностной схемы***), – это число

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i, \quad \text{[бит]}.$$

Замечание. Если $p_i = 0$, то считают $0 \cdot \log_2 0 = 0$.

Подбрасывая монету, закодируем выпадение «орла» – 1, а «решки» – 0. Вероятности этих событий $p(0) = p(1) = 1/2$, распределение вероятностей $\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$. Энтропия, которую получим при подбрасывании монеты:

$$H(X) = -(p_1 \log_2 p_1 + p_1 \log_2 p_1) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = -\log\frac{1}{2} = 1$$

бит.

Введем теперь еще одну дискретную случайную величину Y с распределением вероятностей $P(Y) = \begin{pmatrix} y_1 & y_2 & \dots & y_m \\ p_1 & p_2 & \dots & p_m \end{pmatrix}$ и обозначим

$p(y_j / x_i) = P(\{Y = y_j\} / \{X = x_i\})$ – условную вероятность того, что дискретная случайная величина Y примет значение y_j при условии, что величина X приняла значение x_i. Тогда

- $H(Y / x_i) = -\sum_{j=1}^{m} p(y_j / x_i) \log_2 p(y_j / x_i)$ – условная энтропия случайной величины Y при условии, что величина X приняла значение x_i;
- $H(Y / X) = -\sum_{i=1}^{n} p_i \sum_{j=1}^{m} p(y_j / x_i) \log_2 p(y_j / x_i)$ – полная условная энтропия случайной величины Y.

Свойства энтропии

1. $H(X) \geq 0$, причем $H(X) = 0$, если $p(x_i) = 1$, $p(x_j) = 0$, $i \neq j$;

2. $H(X)$ принимает максимальное значение $H_{\max}(X) = \log_2 n$ (бит), если все n событий равновероятны, т.е. $p_i = \frac{1}{n}$.

3. $H(X,Y) \leq H(X) + H(Y)$, причем равенство, если X, Y – независимы.

4. $H(X,Y) = H(Y) + H(X/Y)$.

5. $H(X/Y) \leq H(X)$, равенство в случае, если X, Y – независимы.

Жизненный опыт подсказывает: часто повторяющиеся события дают нам мало информации. Например, сообщение «11 сентября 2001 г. над Нью-Йорком пролетали самолеты» не вызывает интерес, а известие «11 сентября 2001 г. самолеты разрушили два небоскреба в Нью-Йорке» не оставило никого равнодушным (или сообщение «собака укусила человека» малоинформационно, тогда как сообщение «человек укусил собаку» содержит информации явно больше).

Таким образом, часто повторяющихся и ожидаемые события несут мало информации, а редкие, неожиданные события – больше. Так как информация и вероятность находятся в обратно пропорциональной зависимости, Шеннон предложил формулу:

приращение информации = количеству исчезнувшей неопределенности (энтропии)

Три аксиомы для определения количества информации $I(p)$:

1. Для отдельного события x_i с вероятностью возникновения p_i информация $I(p_i) \geq 0$.
2. Совместная информация двух независимых событий (x_i, x_j) с совместной вероятностью $P(x_i, x_j) = p_{ij} = p_i \cdot p_j$ равна сумме информаций: $I(p_{ij}) = I(p_i) + I(p_j)$.
3. Информация – это непрерывная функция от вероятности события.

По аксиомам 1 и 2 информация нескольких событий не может взаимно уничтожиться.

Количество информации

$$I(X,Y) = H(X) - H(X/Y) \qquad (1)$$

показывает, насколько уменьшается энтропия случайной величины X в результате получения сведений о состоянии случайной величины Y.

§7. ЭНТРОПИЯ ОТКРЫТЫХ ТЕКСТОВ, КЛЮЧЕЙ, ШИФРОТЕКСТОВ

До расшифрования и для законного получателя, и для злоумышленника отправленное сообщение является, естественно, неопределенным. Пусть возможна отправка сообщений $x_1, x_2, ..., x_n$ с вероятностями $p_1, p_2, ..., p_n$ соответственно, т.е. задано распределение вероятностей $P(X)$ открытых текстов. Если какой-либо другой информации о шифровании нет, то остается неопределенность в ответе на вопрос, какой именно из открытых текстов $x_1, x_2, ..., x_n$ шифровался. Мерой такой неопределенности служит энтропия открытых текстов

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i .$$

Она определяет количество битов информации, которое необходимо в среднем передать, чтобы полностью устранить неопределенность в задании сообщения с помощью распределения вероятностей $P(X)$. Другими словами, через эту энтропию определяется количество информации, необходимое для точного указания сообщения.

Если все открытые тексты равновероятны ($p_1 = p_2 = ... = p_n = \frac{1}{n}$), то по свойству 2 энтропия достигает максимального значения $H_{\max} = \log_2 n$. При этом если нет никакой априорной информации о сообщении, кроме его длины в N бит, то все возможные из $n = 2^N$ сообщений считаются равновероятными и тогда энтропия открытого текста равна его длине, поскольку $H_{\max} = \log_2 n = \log_2 2^N = N$.

Если же количество информации о сообщении увеличивается, то распределение вероятностей открытых текстов начинает отличаться от равновероятного, энтропия уменьшается и становится равной нулю, когда есть сообщение x_i, вероятность которого $p(x_i) = 1$ (т.е. $H_{\min} = 0$).

А если об исходном тексте неизвестно вообще ничего, даже его длина? В этом случае все равно необходимо принять за основу какую-либо модель распределения. Как правило, в реальности подобных трудностей не возникает, поскольку большинство шифров не скрывают

размер шифруемого сообщения (учитывая принцип Кирхгоффса). Там же, где этот размер текста необходимо скрыть, все сообщения перед зашифрованием преобразуются в массивы данных одной и той же длины.

Таким образом, энтропия открытого текста измеряет его неопределенность в числе бит информации, которая должна быть восстановлена, когда сообщение было скрыто от криптоаналитика в шифротексте.

После перехвата шифротекста энтропия открытого текста может измениться, теперь она становится апостериорной условной энтропией – условием здесь является перехваченное шифрованное сообщение y:

$$H(X) = -\sum_{i=1}^{n} p(x_i / y)\log_2 p(x_i / y),$$

где $p(x_i / y)$ – вероятность того, что исходное сообщение есть x_i при условии, что результат его зашифрования есть y.

Одной из важнейших характеристик шифра служит количество информации об исходном тексте, которое злоумышленник может извлечь из перехваченного шифротекста – оно находится как разность между априорной и апостериорной энтропией открытого текста:

$$I(X,Y) = H(X) - H(X / Y).$$

Эта величина всегда неотрицательна и показывает, насколько уменьшится неопределенность открытого текста при получении соответствующего шифротекста по сравнению с априорной неопределенностью. Для совершенно стойкого шифра

$$H(X) = H(X / Y),$$

и злоумышленник не может извлечь никакой информации об открытом тексте из перехваченного шифротекста: $I(X,Y) = 0$. Иными словами, знание шифротекста не позволяет уменьшить неопределенность зашифрованного открытого текста и увеличить вероятность его правильного определения.

Аналогично мерами неопределенности шифротекстов и ключей служат энтропии

$$H(Y) = -\sum_{i=1}^{n} p(y_i)\log_2 p(y_i) \text{ и } H(K) = -\sum_{i=1}^{n} p(k_i)\log_2 p(k_i).$$

§8. НЕОПРЕДЕЛЕННОСТЬ ШИФРА ПО КЛЮЧУ

Пусть $H(X), H(Y), H(K)$ – энтропии открытых текстов, шифротекстов и ключей.

По формуле (1)

$I(Y,X) = H(X) - H(X/Y)$ и $I(Y,K) = H(K) - H(K/Y)$

Условную энтропию

$$H(X/Y) = -\sum_{y \in Y} \sum_{x \in X} p(y) \cdot p(x/y) \cdot \log_2 p(x/y) \qquad (2)$$

называют ***неопределенностью шифра по открытому тексту***, а условную энтропию

$$H(K/Y) = -\sum_{y \in Y} \sum_{k \in K} p(y) \cdot p(k/y) \cdot \log_2 p(k/y) \qquad (3)$$

– ***неопределенностью шифра по ключу***. Они измеряют среднее количество информации об открытом тексте и ключе, выдаваемое шифротекстом.

Чем больше $H(X/Y)$, тем меньше информации получит криптоаналитик об открытом тексте по известному шифротексту.

Если для шифра $H(X/Y) = H(X)$, то есть перехваченный шифротекст не дает противнику данных об открытом тексте, то шифр – ***совершенно стойкий.***

Очевидно:

- $H(X/K,Y) = 0$ – по известным криптограмме и ключу мы полностью восстановим открытый текст;
- $H(Y/X,K) = 0$ – зашифровав известный открытый текст на известном ключе, мы получим определенную криптограмму.

Тогда с учетом $p(y/k) = \sum_{\substack{x \in X \\ y = E_k(x)}} p(x)$ и статистически независимых распределений вероятностей открытых текстов и ключей получим

$$H(X,K,Y) = H(X,K) + H(Y/X,K) = H(X,K) = H(X) + H(K);$$
$$H(K,X,Y) = H(K,Y) + H(X/K,Y) = H(K,Y).$$

Левые части последних двух формул одинаковые $\Rightarrow$ равны правые части

$$H(K,Y) = H(X) + H(K).$$

Тогда

$$H(K/Y) = H(K,Y) - H(Y) = H(K) + H(X) - H(Y) \quad \Rightarrow$$

$$\boxed{H(K/Y) = H(K) + H(X) - H(Y)}$$

формула неопределенности шифра по ключу

Пример. Найти энтропию открытых текстов $H(X)$, ключей $H(K)$ и шифротекстов $H(Y)$ для криптосистемы, заданной в примере из §13 (предыдущая лекция). Определить неопределенность шифра по ключу.

Р е ш е н и е. По условию примера из §13 :

$$p(x_1) = 0,24; \quad p(x_2) = 0,16; \quad p(x_3) = 0,28; \quad p(x_4) = 0,32;$$

$$p(k_1) = 0,3; \quad p(k_2) = 0,3; \quad p(k_3) = 0,4;$$

$$p(y_1) = 0,256; \quad p(y_2) = 0,208; \quad p(y_3) = 0,308; \quad p(y_4) = 0,228.$$

$$H(X) = -\sum_{i=1}^{4} p(x_i)\log_2 p(x_i) = -(0,24\log_2 0,24 + 0,16\log_2 0,16 +$$

$$+ 0,28\log_2 0,28 + 0,32\log_2 0,32) \approx 1,956 \text{ (бит);}$$

$$H(K) = -\sum_{i=1}^{3} p(k_i)\log_2 p(k_i) = -(0,3\log_2 0,3 +$$

$$+ 0,3\log_2 0,3 + 0,4\log_2 0,4) \approx 1,570 \text{ (бит);}$$

$$H(Y) = -\sum_{i=1}^{4} p(y_i)\log_2 p(y_i) = -(0,256\log_2 0,256 +$$

$$+ 0,208\log_2 0,208 + 0,308\log_2 0,308 + 0,228\log_2 0,228) \approx 1,985 \text{ (бит).}$$

Таким образом, из шифротекста можно получить $\approx 1,98$ битов информации о ключе и открытом тексте, так как ровно такое количество неопределенности содержит шифротекст.

По формуле для неопределенности шифра по ключу получаем

$$H(K/Y) = 1,570 + 1,956 - 1,985 \approx 1,541 \text{ (бит),}$$

т.е. после перехвата шифротекста остается найти приблизительно $1,5$ битов информации об истинном ключе зашифрования. Пример демонстрирует, почему, такая криптосистема не надежна: если вначале было $1,570$ бит неопределенности относительно ключа, а перехват шифротекста уменьшил неопределенность до $1,541$ бит, то шифротекст «разглашает» о ключе $1,570 - 1,541 = 0,029$ (бит) информации.

§9. ЭНТРОПИЯ И ИЗБЫТОЧНОСТЬ ЯЗЫКА

Энтропия языка – это количество информации, которая приходится на одну букву открытого осмысленного текста на этом языке. Она отображает вероятностно-лингвистические связи текста. В теории информации доказано, энтропия языка совпадает с условной энтропией буквы открытого текста, если известны все буквы, стоящие перед ней. Энтропию языка оценивают с помощью последовательных приближений:

1) I приближение энтропии языка с алфавитом из m букв – это ***абсолютная энтропию языка*** $H_0 = \log_2 m$. (*то есть максимальная энтропия отдельных независимых букв*).

- для языков с латинским алфавитом (26 букв и пропуск)
$$H_0 = \log_2 27 \approx 4,76 \text{ битов.}$$
- для русского языка $H_0 = \log_2 32 \approx 5$ битов,
- для украинского языка $H_0 = \log_2 34 \approx 5,09$ битов.

2) II приближение энтропии языка – это $H_1 = -\sum_{i=1}^{m} p_i \log_2 p_i$, где p_i - вероятность повторения букв алфавита в языке (*энтропия распределения букв алфавита*).

3) III приближение энтропии языка – это энтропия H_2 распределения биграмм, деленная на 2 (нас интересует энтропия на одну букву).

4) IV приближение энтропии языка – это энтропия H_3 распределения триграмм, деленная на 3.

Обобщим. ***Энтропией языка*** $H_{\text{яз.}}$ называют предел

$$\lim_{r\to\infty} \frac{H_r}{r} = H_{\text{яз.}},$$

где H_r – энтропия распределения r – грам.

Величину

$$D = 1 - \frac{H_{\text{яз.}}}{\log_2 m},$$

соответственно называют ***избыточностью языка*** (иногда ее выражают в процентах). Она показывает, какую часть букв открытого текста можно выбросить без потери смысла (*то есть потерянные буквы и вся информация можно восстановить через другие буквы текста с помощью лингвистических связей*).

Ошибка считать, что если избыточность языка 75 %, то вычеркнув любые три из каждых 4 букв текста, можно однозначно восстановить исходную информацию. Избыточность языка характеризует неиспользованные возможности при «сжатии» данных в открытому тексте. Так, при оптимальном кодировании текста, например, кодами Хоффмана, Фано, написанного языком с избыточностью 75 %, текст можно «сжать» на 3/4 его длины без потери информации.

Энтропия и избыточность языков

		Английский	Русский	Украинский
Энтропия (бит/букву)	H_0	4,76	5	5,09
	H_1	4,03	4,35	4,51
	H_2	3,56	3,52	3,47
	H_3	3,30	3,01	3,14
	H_9			1,4
	$H_{\text{яз.}}$	1,0 – 1,5	1,19 –1,40	1,25 –1,40
Избыточность D, (%)		68,0–72,3	72,0 – 76,2	72,5 – 75,4

Замечание 1. Для всех языков $H_N^{\text{с пропус.}} < H_N^{\text{без пропус.}}$.

§10. РАССТОЯНИЕ ЕДИНСТВЕННОСТИ

Зашифруем с помощью шифра сдвига $y = x + k (\mathrm{mod}\, 33)$ (укр.яз):

МАСА с ключом $k = 2 \Rightarrow 18\ \ 02\ \ 23\ \ 02 \Rightarrow$ ОВУВ

ТЕЧЕ с ключом $k = 29 \Rightarrow 18\ \ 02\ \ 23\ \ 02 \Rightarrow$ ОВУВ

Криптограммы совпали. Подобная ситуация возможна для любого шифра. При попытке найти истинный ключ шифра по данной криптограмме полным перебором ключей криптоаналитик может получить несколько осмысленных открытых текстов для разных ключей. При этом только один из ключей будет ***истинным ключом***, а остальные – ***фальшивыми***.

Пусть открытые тексты осмыслены, имеют длину L, записаны в алфавите A из m букв природным языком с избыточностью D. Обозначим

K_y – множество тех ключей, с помощью которых при потоковом шифровании некоторого осмысленного открытого текста появится одна и та же криптограмма y. Очевидно, только один ключ из множества ключей K_y будет истинным. Справедлива теорема.

Теорема об оценке среднего числа фальшивых ключей. Для потокового шифра с равновероятным выбором ключей при достаточно больших значениях L среднее число k_L фальшивых ключей связано с избыточностью языка D :

$$k_L \geq \frac{\left|K_y\right|}{m^{L \cdot D}} - 1,$$

де $\left|K_y\right|$ – количество ключей в множестве K_y.

Атакуя шифр, хотелось бы снизить количество фальшивых ключей до нуля. При росте длины шифрованного текста количество фальшивых ключей уменьшается. ***Расстояние единственности шифра по ключу*** – это минимальная длина L_0 шифрованного текста, необходимого для однозначного восстановления истинного ключа шифра.

Из теоремы об оценке среднего числа фальшивых ключей

$$\frac{\left|K\right|}{k_L + 1} \leq m^{L \cdot D}.$$

Если $k_L = 0$, то $\left|K\right| \leq m^{L \cdot D}$, откуда

$$L \geq \frac{\log_2 \left|K\right|}{D \cdot \log_2 m}. \qquad (4)$$

Наименьшее целое число L_0, удовлетворяющее неравенству, принимается за расстояние единственности (m – количество букв алфавита; $\left|K\right|$ – количество ключей шифра; D – избыточность языка). В идеально стойких шифрах $L_0 = \infty$, но для других шифров расстояние единственности может быть опасно малым.

Аналогично вычисляют ***расстояние единственности шифра по открытому тексту*** (минимальное длина L_0 криптограммы, по которой однозначно восстанавливается открытый текст).

Пример. В открытых текстах использовано 33 буквы + пропуск. Избыточность языка 72 %. Найти расстояние единственности шифра моноалфавитной подстановки.

Р е ш е н и е. $m = 34$; $D = 0,72$. Шифр имеет $m! = 34!$ разных ключей. По формуле Стирлинга $m! \approx \left(\frac{m}{e}\right)^m \cdot \sqrt{2\pi m}$ имеем

$$\ln(m!) \approx \left(m + \frac{1}{2}\right) \cdot \ln m - m + \frac{1}{2}\ln(2\pi).$$

$$L \geq \frac{\log_2 |K|}{D \cdot \log_2 m} = \frac{\ln |K|}{D \cdot \ln m} = \frac{\ln(34!)}{0,72 \cdot \ln 34} \approx \frac{34,5 \cdot \ln 34 - 34 + 0,5 \ln(6,28)}{0,72 \cdot \ln 34},$$

$$L \geq 34,88 \ ,$$

расстояние единственности шифра $L_0 \approx 35$.

Шифруя с помощью моноалфавитной подстановки открытый текст длиной 35 букв, при перехвате криптограммы можно ожидать, что текст будет однозначно восстановлен.

Если перед передачей текст сжимается и употребляется идеально архивирующий алгоритм, то сжатый текст уже не будет избыточным и $D \to 0$, а расстояние единственности

$$L_0 \approx \frac{l}{0} = \infty .$$

Криптоаналитики используют естественную избыточность языка для уменьшения числа вероятных открытых текстов. Чем избыточнее язык, тем легче его криптоанализировать. По этой причине часто **перед** шифрованием используют программы сжатия для уменьшения размера текста. Сжатие уменьшает избыточность сообщения вместе с объемом работы, необходимым для его шифрования и дешифрирования.

Таким образом, шифротексты, которые длиннее расстояния единственности, можно дешифровать, вероятнее всего, только одним осмысленным способом. Что касается шифротекстов, которые заметно короче расстояния единственности, их можно дешифровать, скорее всего, несколькими способами, каждый из которых может быть правилен, и таким образом обеспечить безопасность, поставив перед противником задачу выбора правильного открытого текста.

Расстояние единственности является не точным, а вероятностным значением. Оно позволяет оценить минимальное количество шифротекста, при вскрытии которого с помощью метода полного перебора ключей, вероятно, появится только один разумный способ дешифрирования. Обычно чем больше расстояние единственности, тем более стойкая криптосистема.