ЭНТРОПИЯ ПОЛЬЗОВАТЕЛЬСКИХ ПАРОЛЕЙ

Л.А. Савостин

Белорусский государственный университет информатики и радиоэлектроники
кафедра программного обеспечения информационных технологий
г. Минск, Беларусь

Рассматривается зависимость надежности пароля от последовательности символов. Приводятся примеры частотного распределения символов английского алфавита и цифр. Показывается зависимость сложности пароля от его длины и объема алфавита. Приводятся рекомендации для улучшения надежности паролей.

Ключевые слова – энтропия, пароль, безопасность, защита информации.

1 ОСНОВНАЯ ЧАСТЬ

На практике часто используются легко запоминающиеся уязвимые пароли. Такие пароли обычно являются словами, выражениями, аббревиатурами или иными языковыми конструкциями, принадлежим естественным языкам. Пароли, обладающие вышеописанными свойствами, очень уязвимы и легко поддаются атакам. Использование надежных паролей, как правило, вызывает у пользователей неудобства, поскольку они сложны для запоминания. Надежные с точки зрения безопасности пароли представляют собой длинные последовательности, состоящие из большого разнообразия символов.

Методы криптоанализа позволяют произвести атаку на пароль, последовательность символов которого обладает частотными свойствами. Легко запоминающиеся пароли обладают такими свойствами, постольку распределение букв естественного языка также обладает частотными свойствами.

Расширение алфавита используемого для представления пароля (добавления буквы, цифры и специальных символов), приводит к увеличению надежности. При использовании последовательности, не являющейся корректной с грамматической точки зрения, надежность также увеличивается. И наоборот, запоминающиеся обрывки слов или фраз, приводят к снижению стойкости. Йоханссон приводит такие методы как умышленное добавление символов, что эффективно увеличивает стойкость фразы или пароля. Сох предлагает использовать бессмысленные выражения вместе с фразой или предложением, которые легко запомнить, в качестве повышения надежности.

При нынешних компьютерных мощностях можно легко перебрать все перестановки коротких паролей, состоящих из символов, находящихся на клавиатуре. Бернетт утверждает, что безопасный пароль должен состоять минимум из двадцати символов [1 с.121-4].

Надежность пароля увеличивается при увеличении следующих его параметров: объема алфавита и длины. Вместе эти параметры определяют энтропию, или вероятность распределения пароля. Таблицы 1, 2 и 3 отображают зависимость времени атаки от размера алфавита и длинны пароля. Энтропия – это мера беспорядка в системе. В информационных системах она может рассматриваться как мера недостатка информации в последовательностях. Шеннон показал, что энтропия дискретной случайной величины x, из набора n выражается формулой ниже:

$$H(x) = \sum_{i=1}^{n} p(i)\log_2\left(\frac{1}{p(i)}\right) = -\sum_{i=1}^{n} p(i)\log_2 p(i) \qquad (1.1)$$

Энтропия события x является суммой с противоположным знаком всех произведений относительных частот появления события i, умноженных на их же двоичные логарифмы [2]. Основание 2 выбрано только для удобства работы с информацией, представленной в двоичной форме. Энтропия каждого случайного символа в текстовой строке является двоичным логарифмом ряда вероятностей и таким образом энтропия всей строки зависит от энтропии каждого символа. Случайность в наборе символов и последовательности символов определяет энтропию пароля, таким образом, энтропия является прямым показателем надежности пароля.

Удобство паролей зависит от легкости ввода и запоминания. Обычно удобные пароли являются частью естественного языка. В работе рассматриваются примеры на английском, но естественные языки имеют схожие структуры. Частотное распределение букв английского алфавита выглядит следующим образом: A-7.3, N-7.8, B-0.9, O-7.4, C-3.0, P-2.7, D-4.4, Q-0.3, E-13.0, R-7.7, F-2.8, S-6.3, G-1.6, T-9.3, H3.5, U-2.7, I-7.4, V-1.3, J-0.2, W-1.6, K-0.3, X-0.5, L-3.5, Y-1.9, M-2.5, Z-0.1. Легко заметить что гласные «E», «I», «O», «A» и «U» являются наиболее встречаемыми буквами, и соответственно им согласные «T», «N», «R», «D» и «L». Буквы «K», «Q», «X», «J» и «Z» - встречаются реже всех. Пять гласных букв составляют приблизительно 40% английского текста, пять согласных («D», «N», «R», «S» и «T») составляют 35%, десять среднечастотных согласных («B», «C», «F», «G», «H», «L», «M», «P», «V» и «W») - лишь 24%, а пять согласных низкой частоты («J», «K», «Q» и «Z») составля-

ют лишь 1%[3].

Распределения любого текста, например пароль «Beowulf», или сообщения передаваемого по электронной почте могут быть различными. Если пароль был зашифрован подстановочным или сдвиговым шифром, то полученная зашифрованная последовательность также будет обладать частотными свойствами, которым возможно статистически проанализировать.

Цифры также обладают частотными характеристиками (%): «1»-21, «2»-13, «3»-9, «4»-8, «5»-8, «6»-8, «7»-8, «8»-7, «9»-9, «0»-10. Цифра «1» выделяется, этот факт увеличивает шансы злоумышленников. В дополнение к одиночным символам, существую и другие шаблоны в естественных языках, имеющие статистические характеристики. Ниже приведены двуграммы английского языка с высокой частотой появления. Для опыта использовались последовательности в 200 символов:TH-50, ER-40, ON-39, AN-38, RE-36, HE-33, IN-31, ED-30, ND-30, HA-26, AT-25, EN-25, ES-25, OF-25, OR-25, NT-24, EA-22, TI-22, TO-22, IT-20, ST-20, IO-18, LE-18, IS-17, OU-17, AR-16, AS-16, DE-16, RT-16 и VE-16.

Наиболее частыми триграммами в английском языке являются (для опыта использовалась последовательности в 200 символов): THE-89, AND-54, THA-47, ENT-39, ION-36, TIO-33, FOR-33, NDE-31, HAS-28, ACE-27, EDT-27, TIS-25, OFT-23, STH-21 и MEN-20.

Распределение наблюдается также на границах слов в естественных языках. Если злоумышленник сможет определить границы слов, то у него появится возможность сократить пространство поиска.

Английский язык имеет примерно 75% избыточности[4, с. 64], таким образом, энтропия пароля состоящего из английских слов составляет только 25% от энтропии случайной последовательности. Предсказуемость слов английского языка подтверждает тот факт, что существуют лишь 17000 слов, состоящих из восьми букв, из которых только 500 находятся в общем использовании, хотя имеются 208,827,064,576 доступных комбинаций из восьми символов[1]. Отсутствие случайности становится более очевидным, когда пользователь выбирает пароля. Хорошие пароли могут частично компенсировать этот недостаток наличием случайных чисел или символов.

Еще одним средством для улучшения энтропии является перестановка, которая представляет ряд возможных комбинаций. Количество возможных комбинаций можно определить формулой 1.2:

$$P(n,r) = \frac{n!}{(n-r)!}. \qquad (1.2)$$

где n общее количество объектов для выбора и r – число объектов, которые будут выбраны[4]. Таким образом, если клавиатура содержит 95 клавиша, то количество возможных восьми символьных паролей равно 95!/(95-8)!. Полностью случайный восьми символьный пароль, состоящий из комбинации 95 возможных ASCII символов, как правило, содержит более 52 бит энтропии, в то время восьми символьный пароль из букв одного регистра и цифр содержит 41 бит.

Энтропия в рамках любого текста это всегда добавление. Поэтому вместе с пространством поиска, дополнительная энтропия требует дополнительных нажатий на клавиши, и при равных других условиях, более длинные пароли содержат больше энтропии, чем короткие.

Эксперты в области безопасности и администраторы рекомендуют использовать большой алфавит для увеличения энтропии и сложности атаки. В случае американской клавиатуры, предел алфавита составляет 95 ASCII символов. Но при желании можно также использовать символы Unicode, путем ввода 4-значного номера, удерживая клавишу Alt.

В таблице 1 показано отношение пространства поиска и энтропии паролей. Например, при алфавите в 26 букв энтропия составляет 4,7 бита на букву.

Таблица 1

Отношение пространства поиска и энтропии

Пространство поиска	N	Энтропия, бит/символ
Только цифры (0-9)	*10*	*3.3*
Буквы нижнего регистра (a-z)	*26*	*4.70*
Буквы нижнего регистра и цифры (a-z, 0-9)	*36*	*5.17*
Буквы обоих регистров и цифры (a-z, A-Z, 0-9)	*62*	*5.9*
Весь набор клавиш, доступных на клавиатуре	*94*	*6.55*
Unicode символы	*700*	*9.50*

Исследователи установили, что энтропия для пароля из пространства 700 Unicode символов будет 9,5 бита на символ, и посчитали, что каждое последующее удвоение пространства поиска увеличит энтропию на один бит.

ЛИТЕРАТУРА

[1] Burnett, M. Perfect Passwords: Selection, Protection, Authentication. Rockland / Burnett, M. – MA: Syngress, 2004. p 133.

[2] Shannon, C. A Mathematical Theory of Communication / Shannon, C. - Bell System Technical Journal,27, 1948. - p. 379-423, 623-656.

[3] Friedman, W. The index of coincidence and its applications in cryptanalysis.Technical Paper, War Department, Office of the Chief Signal Officer / Friedman, W. - Washington: United States Government Printing Office, 1925. - p. 52.

[4] Stinson, D. Cryptography: Theory and Practice / Stinson, D. Second Edition. Boca Raton, FL: Chapman & Hall/CRC, 2002. – p. 192.

[5] Knuth, D. The Art of Computer Programming, Volume 3: Sorting and Searching, / Knuth, D. Third Edition. New York: Addison-Wesley, 1997. p. 123.