УДК 004.33.054

А. А. Савостин

ЭНТРОПИЯ ПАРОЛЕЙ

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.

Введение

На практике часто используются легко запоминающиеся уязвимые пароли. Такие пароли обычно являются словами, выражениями, аббревиатурами или иными языковыми конструкциями, принадлежащим 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 [1]. Кох предлагает использовать бессмысленные выражения вместе с фразой или предложением, которые легко запомнить, в качестве повышения надежности [2].

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 [3, с. 121–124]. Использовать длинные пароли также рекомендуют Портер [4] и Плиджер [5].

Энтропия паролей

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 длины 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:

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

Энтропия события x является суммой с противоположным знаком всех произведений относительных частот появления события i, умноженных на их же двоичные логарифмы [6]. 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 изображено в табл. 1.

Таблица 1

Частотное распределение букв английского алфавита

A	B	C	D	E	F	G	H	I	J	K	L	M
7.3	0.9	3.0	4.4	13.0	2.8	1.6	3.5	7.4	0.2	0.3	3.5	2 5
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
7.8	7.4	2.7	0.3	7.7	6.3	9.3	2.7	1.3	1.6	0.5	1.9	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») – 24 %, а пять согласных низкой частоты («J», «K», «Q» и «Z») составляют 1 % [7]

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. В табл 2 приведены примеры.

Таблица 2

Частотное распределение цифр

0	1	2	3	4	5	6	7	8	9
10 %	21 %	13 %	9 %	8 %	8 %	8 %	8 %	7 %	9 %

Из таблицы 2 видно, что цифра «1» выделяется, этот факт увеличивает шансы злоумышленников. В дополнение к одиночным символам, существуют 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 [8].

Распределение наблюдается также на границах слов в естественных языках. Например, частоты первых и последних букв английских слов приведены в табл 3.

Таблица 3

Распределение букв на границах слов в английском языке

Буква	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Первая	9	6	–	5	2	4	2	3	3	1	1	2	4	2	10	2	–	4	5	17	2	–	7	–	3	–
Последняя	1	–	10	1	7	6	4	2	–	–	1	6	1	9	4	1	–	8	9	11	1	–	1	–	8	

Если злоумышленник сможет определить границы слов, то у него появится возможность сократить пространство поиска.

Английский язык имеет примерно 75 % избыточности [8, 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 17 000 слов, состоящих из восьми букв, из которых только 500 находятся в общем использовании, хотя имеются 208, 827, 064, 576 доступных комбинаций из восьми символов[3]. 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 2:

$$P(n,r)=\frac{n!}{(n-r)!}, \quad (2)$$

где n общее количество объектов для выбора и r – число объектов, которые будут выбраны [9]. Таким образом, если клавиатура содержит 95 клавиш, то количество возможных восьмисимвольных паролей равно 95!/(95-8)! Полностью случайный восьмисимвольный пароль, состоящий из комбинации 95 возможных ASCII символов, как правило, содержит более 52 бит энтропии, в то время восьмисимвольный 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 четырехзначного номера, удерживая клавишу Alt.

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

Таблица 4

Отношение пространства поиска и энтропии

Пространство поиска	N	Энтропия
Только цифры (0–9)	10	3.32 бит/символ
Буквы нижнего регистра (a–z)	26	4.70 бит/символ
Буквы нижнего регистра и цифры (a–z, 0–9)	36	5.17 бит/символ
Буквы обоих регистров и цифры (a–z, A–Z, 0–9)	62	5.95 бит/символ
Весь набор клавиш, доступных на клавиатуре	94	6.55 бит/символ
Unicode символы	700	9.50 бит/символ

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.

Заключение

При увеличении энтропии возможно добиться улучшения надежности парол, являющегося элементом естественного языка. Она не является фиксированной физической величиной. Криптоаналитики могут использовать шаблоны языка, распределение частот символов, биграммы, чтобы сократить пространство поиска. Факт, что при увеличении длины пароля энтропия тоже увеличивается, является особенно важным для коротких паролей.

Литература

1. The Great Debates: Pass Phrases vs. Passwords. Part 3 of 3.[Electronic resource} / Johansson, J. 1996. – Mode of access: http://www.microsoft.com/ technet/security/

2. B. PGP Passphrase FAQ. FAQ: How do I choose a good password or phrase?/ [Electronic resource] / Cox – Mode of access http://www.virtualschool.edu/ mon/Crypto/PGPPassPhraseFAQ.html.

3. *Burnett, M.* Perfect Passwords: Selection, Protection, Authentication. Rockland / M. Burnett. – MA: Syngress, 2004. – P 133.

4. *Porter, S.* A Password Extension for Improved Human Factors. In A. Gersho / S. Porter. – Computers & Security, 1982. – P. 54.

5. *Pleeger, C.* Security in Computing. Englewood Cliffs / C. Pleeger. – NJ: Prentice.

6. Hall, 1989. – P. 23.

7. *Shannon, C.* A Mathematical Theory of Communication / C. Shannon. – Bell System Technical Journal, 27, 1948. – P. 379–423, 623–656.

8. *Friedman, W.* The index of coincidence and its applications in cryptanalysis.Technical Paper, War Department, Office of the Chief Signal Officer / W. Friedman. – Washington: United States Government Printing Office, 1925. – P. 52.

9. *Stinson, D.* Cryptography: Theory and Practice / D. Stinson. Second Edition. Boca Raton, FL: Chapman & Hall/CRC, 2002. – P. 192.

10. *Knuth, D.* The Art of Computer Programming, Volume 3: Sorting and Searching, / D. Knuth, Third Edition. New York : Addison-Wesley, 1997. – P. 123.

Савостин Артем Алексеевич, аспирант кафедры программного обеспечения информационных технологий Белорусского государственного университета информатики и радиоэлектроники, artem.savostin@gmail.com

УДК 515.142.33

М. В. Стержанов

АЛГОРИТМ ПОСТРОЕНИЯ ГРАФОВОЙ И ВЕКТОРНОЙ МОДЕЛЕЙ БИНАРНОГО РАСТРА, КОДИРОВАННОГО ДЛИНАМИ СЕРИЙ

Векторизация – это процесс построения векторной модели растрового изображения. Для более точной обработки важно иметь промежуточное представление. Мы предлагаем строить графовую модель. Она сохраняет топологию и используется в процедуре векторизации. Мы предлагаем алгоритмы для построения графовой и векторной моделей.

Введение

Векторизация -- это процесс построения векторной модели растрового изображения. Растровое изображение представляет собой прямоугольную матрицу, значение элемента которой (пикселя) соответствует его яркости. В бинарном изображении каждый пиксель окрашен в белый (фон), либо в черный (изображение) цвет. Под векторной формой понимается цифровое представление точечных, линейных и полигональных пространственных объектов в виде набора координатных пар, с описанием только геометрии объектов. Как следствие, векторная форма хранения характеризуется компактностью, простотой обращения к объектам, с ее помощью несложно осуществлять трехмерное моделирование объектов.

При сканировании изображения с высоким разрешением выходной файл получается достаточно большого размера. Это накладывает ограничения на применяемые алгоритмы обработки. Необходима структура данных или другими словами представление, которая будет обеспечивать компактное хранение изображения, сохраняя его топологию. Желательно, чтобы данное представление хранило и морфологические свойства изображения. Рассмотрим сильные и слабые стороны некоторых методов представления растровой информации. Операция утоньшения позволяет представить объекты на растре линиями единичной ширины. Это дает возможность эффективно кодировать объекты и формировать цифровую модель. Операция утоньшения может применяться в качестве операции сортировки по величине, например для выделения на изображении наибольшего связного замкнутого объекта. Скелетизированное изображение (СИ) сохраняет топологию, однако оно чувствительно к шуму, места соединений обрабатываются не всегда корректно. Некоторые алгоритмы скелетизации не сохраняют связность. Площадные объекты в СИ представляются искаженно. Алгоритмы выделения контуров можно условно разбить на две группы: отслеживающие и сканирующие. Недостатком контурного препарата является то, что по нему трудно построить топологию исходного изображения. Граф смежности линий является удобным способом представления изображения, состоящего из большого числа горизонтальных или вертикальных отрезков. Однако данная структура не хранит информацию о местах соединения, что, безусловно, является серьезным недостатком.

Предлагаемый метод векторизации является гибридным, т. е. сочетает в себе различные подходы. Для ускорения выделения связных компонент и скелетизации используется кодирование длинами серий. Затем, анализируя связность смежных серий, выделяются полосы. Средние линии полос представляются скелетными кривыми. Зоны соединений выделяются, рассчитывается центр пересечения, затем выполняется попиксельное утоньшение. Строится графовая структура, описывающая топологию СК. Ребрам графа присваиваются весовые параметры, отражающие толщину, площадь и другие морфологические признаки компонент изображения. Ребра графа описывают «направления» векторизации. Применяется алгоритм генерализации, выделяющий отрезки. Затем находятся дуги окружностей.