Cripto-análisis sobre métodos clásicos de cifrado

Crypto-analysis over classical encryption methods

Sebastián Gómez, Juan David Arias, Diego Agudelo
Ingeniería de Sistemas, Universidad Tecnológica de Pereira, Pereira, Colombia
segomez@utp.edu.co
judavid.arias@gmail.com
dialagudelo@utp.edu.co

***Resumen*— Este artículo muestra algunas técnicas de encripción clásicas como los cifrados del Cesar y Vigenère. Este artículo también muestra algunas técnicas básicas y modernas de cripto-análisis basadas en la teoría de la información y la estadística. Estas técnicas de cripto-análisis también pueden ser usadas en otros métodos de cifrado si las mismas debilidades están presentes. Por esta razón es útil conocer estas técnicas de cripto-análisis cuando se diseñan nuevos algoritmos.**

***Palabras clave*— cifrado del cesar, cifrado de Vigenère, criptoanálisis, criptografía, entropía, estadística, teoría de la información.**

***Abstract*— This article shows some classic encryption methods like Caesar and Vigenère ciphers. This article also shows some basic and modern crypto-analysis techniques based on information theory and statistics. These crypto-analysis techniques can also be used on other encryption methods if the same weaknesses are present. For this reason it is useful to know these crypto-analysis techniques when designing new cryptographic algorithms.**

***Key Word* — Caesar cipher, Vigenère cipher,crypto-analysis, cryptography, entropy, statistics, information theory .**

I. INTRODUCCIÓN

Los algoritmos criptográficos utilizados desde la época de la antigua Roma hasta nuestros días, son métodos que convierten un mensaje de texto plano en texto cifrado. El proceso inverso, se conoce como "descifrar", y consiste en llevar el texto cifrado a texto claro. Usualmente estos algoritmos utilizan una llave secreta como parte de la entrada. En los métodos que se van a mostrar en este artículo, conocidos como cifrado de llave simétrica, la llave debe coincidir tanto en el proceso de cifrado como en el de descifrado para que la comunicación entre el emisor quien cifra, como el receptor que descifra, sea exitosa.

El objetivo de realizar el cifrado, es dificultar (o imposibilitar) la comprensión del mensaje a otra persona distinta del receptor legitimo del mensaje. Idealmente los emisores y receptores legítimos del mensaje deben ser los únicos que conozcan la llave secreta, ya que es en esta llave secreta que radica la privacidad de la comunicación. Cabe recalcar que la seguridad de la información no solo radica en la privacidad, sino también en la integridad y en algunos otros factores importantes. Sin embargo en este artículo el cifrado se enfoca en el aspecto de la privacidad.

Se conoce como criptoanalista a la persona que sin tener la llave secreta, trata de descifrar un texto encriptado. Los estándares criptográficos actuales exigen que aunque el algoritmo de encripción sea conocido por el criptoanalista, si este no posee la llave secreta, no sea factible que este consiga obtener todo o parte del texto plano. Es por este motivo que al diseñar algoritmos criptográficos modernos se deben conocer los métodos de criptoanálisis. El conocimiento de métodos de criptoanálisis permite al diseñador de algoritmos criptográficos analizar que debilidades y fortalezas tienen los algoritmos, y de esta forma estar mejorándolos continuamente. [1]

En este artículo se pretenden describir algunas de las técnicas de cifrado clásicas y mostrar los conceptos y herramientas necesarias para hacer el respectivo criptoanálisis de dichos métodos.

II. CIFRADO DEL CESAR

El cifrado del cesar es uno de los más simples, usado por el cesar para comunicarse con sus generales, es un tipo de cifrado por sustitución en el que un símbolo del texto plano es sustituido por otro símbolo que se encuentra k posiciones adelante en el alfabeto. Por ejemplo si suponemos que el alfabeto es el alfabeto del español y tenemos como texto plano "holaquetal", y suponemos que k = 1 entonces cada letra se reemplazaría por la siguiente en el alfabeto. De esta forma la 'h' se convertiría en 'i', la 'o' en 'p' y asi sucesivamente hasta obtener el texto encriptado "ipmbrvfubm". Formalmente se puede definir el cifrado del Cesar como se muestra en la ecuación (1):

$$C_i \equiv S_i + K \ (mod\ N) \qquad (1)$$

Donde S_i corresponde al carácter en la posición i del texto plano, C_i corresponde al carácter i del texto encriptado, y N corresponde a la cantidad de símbolos del alfabeto. La desencripción sería de una forma similar.

Fecha de Recepción: 26 de Agosto de 2011
Fecha de Aceptación: 30 de Abril de 2011

$$C_i \equiv S_i - K \ (mod\ N) \qquad (2)$$

El problema evidente que tiene este método es que un mismo símbolo en el texto plano, se encripta al mismo símbolo en el texto encriptado. De esta forma si un criptoanalista sabe que se utilizó este método podría simplemente ver cuál es la letra más repetida en el texto encriptado y sabiendo que en el alfabeto español la letra 'e' es la que más se repite podría simplemente hacer la resta entre estas dos letras y hallar k fácilmente [2].

III. CIFRADO DE VIGENÉRE

El cifrado de Vigenère está basado en el cifrado del Cesar [2], por lo cual es un cifrado de sustitución. A diferencia del cifrado del cesar, en el cual cada símbolo del texto plano le es sumada una constante k, en el cifrado de Vigenère se tiene un cifrado del Cesar por cada símbolo de una palabra clave. Con lo cual si la palabra clave tiene una longitud m, se tienen m corrimientos diferentes sobre el texto encriptado. De esta forma, no siempre un mismo símbolo en el texto claro se convierte en el mismo símbolo en el texto encriptado [3].

A	0	J	9	R	18
B	1	K	10	S	19
C	2	L	11	T	20
D	3	M	12	U	21
E	4	N	13	V	22
F	5	Ñ	14	W	23
G	6	O	15	X	24
H	7	P	16	Y	25
I	8	Q	17	Z	26

Tabla 1.

Si cada símbolo del alfabeto representara un número del 0 al 26, como se muestra en la Tabla 1, se seguiría el siguiente procedimiento: A cada letra del texto plano se le sumaría una letra de la clave y como la clave suele ser de menor longitud que el texto plano se repetiría para lograr el tamaño del texto plano. Un ejemplo de esto se muestra en la Tabla 2.

H	O	L	A	Q	U	E	T	A	L
K	E	Y	K	E	Y	K	E	Y	K
Q	S	J	K	U	S	Ñ	X	Y	U

Tabla 2. Ejemplo de cifrado.

En el ejemplo de la tabla 2, se muestra en la primera fila el texto plano a cifrar, en la siguiente fila la clave de cifrado "KEY" repetida hasta lograr la longitud del texto plano y en la última fila se muestra el texto cifrado. Para hacer el cifrado la letra 'H' (7) se le suma la letra 'K' (10) y se obtiene 'Q' (17), a la 'O' (15) se le suma la 'E' (4) y se obtiene la 'S' (19), a la 'L' (11) se le suma la 'Y' (25) y se obtiene la 'J' ($11 + 25 \equiv 36 \equiv 9\ (mod\ 27)$) y así sucesivamente.

Formalmente el cifrado de Vigenère se puede expresar de la siguiente manera. Suponga que n representa la cantidad de símbolos del alfabeto, m representa la longitud de la clave, S_i corresponde al carácter en la posición i del texto plano, K_i corresponde al carácter i de la palabra clave, y C_i corresponde al carácter i del texto encriptado. Entonces:

$$C_i \equiv S_i + K_{i\ mod\ m}\ (mod\ n) \qquad (3)$$

Para la desencripción el texto conociendo la clave:

$$C_i \equiv S_i - K_{i\ mod\ m}\ (mod\ n) \qquad (4)$$

IV. ENTROPIA

La entropía es un concepto valioso cuando se piensa en hacer criptoanálisis dado que representa la medida promedio de información que tiene un símbolo en algún mensaje, de hecho se puede pensar en calcular la entropía para cierto lenguaje (español, inglés, etc.) y es curioso saber que la entropía de cada lenguaje tiende a cierto valor característico. [4]

La cantidad de información de un símbolo B se define como:

$$I(B) = log_2(\frac{1}{P(B)}) \qquad (5)$$

Donde $P(B)$ representa la probabilidad de aparición en el mensaje del símbolo B, esta expresión tiene sentido si se piensa en el hecho de que entre más alta sea la probabilidad de aparición de un símbolo, menos cantidad de información representa para el mensaje, y entre más improbable sea la aparición del símbolo, más información representa para el mensaje, como se puede ver en la figura 1:

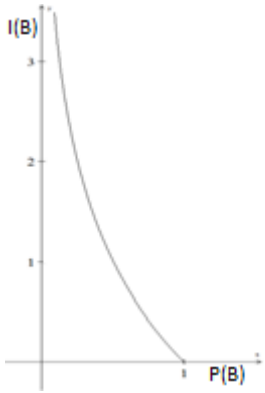

Figura 1

Teniendo en cuenta lo anterior la entropía para un mensaje M se define así

$$H(M) = \sum_{i=0}^{N-1} P(T_i) * I(T_i) \qquad (6)$$

Donde T_i es cada uno de los símbolos que componen el alfabeto del mensaje M, I es la función de cantidad de información definida en (5) y N es el número de símbolos del alfabeto. Se puede ver que $H(M)$ representa el valor medio de la cantidad de información de cada símbolo del alfabeto dadas las probabilidades de aparición de estos, por lo tanto H(M) representa la cantidad de información que se esperaría encontrar para un símbolo del mensaje M.

Para ilustrar la utilidad de la entropía en el criptoanálisis se puede tomar como ejemplo el cifrado de Vigenère descrito con anterioridad, supongamos que se tiene un texto encriptado y que se conoce que fue encriptado con el método de Vigenère. Además de esto se conocen las probabilidades de aparición de los símbolos que constituyen el lenguaje original del mensaje encriptado, pero se desconoce la clave con la que se hizo el cifrado. Un posible ataque para el método de Vigenère con las condiciones dadas puede partir de la siguiente observación.

Si el mensaje original fue encriptado con una clave de longitud T, y se toman símbolos cada T caracteres del texto cifrado estos debieron haber sido sometidos a la misma operación de cifrado, es decir:

El conjunto de símbolos C_j tal que $C_j = S_{T*i+j}$ para todo i tal que $i \times T < |M|$ fueron sometidos a la misma transposición que denota el carácter j-ésimo de la clave, por lo tanto cada uno de estos conjuntos C_j se puede ver como un cifrado del Cesar, de este modo el primer problema que se debe enfrentar es el de averiguar la longitud de la clave de encripción para posteriormente atacar cada uno de los cifrados del Cesar por un ataque de máxima correlación (Sección 5).

Para el ejemplo de tabla 2 se puede observar que se tienen 3 conjuntos C_j (C0, C1, C2), cada uno de estos conjuntos fue sometido a la misma transposición por lo que cada uno es un cifrado del Cesar independiente, así las letras del conjunto C0 fueron cifradas con el carácter 'K', el conjunto C1 fue cifrado con el carácter 'E' y el conjunto C2 fue cifrado con el caracter 'Y', esto se muestra en tabla 3.

H	O	L	A	Q	U	E	T	A	L
K	E	Y	K	E	Y	K	E	Y	K
Q	S	J	K	U	S	Ñ	X	Y	U

	Letras del conjunto C0
	Letras del conjunto C1
	Letras del conjunto C2

Tabla 3

Para averiguar la longitud de la clave se tiene en cuenta el hecho de que generalmente la clave es pequeña en comparación con el mensaje que se encripta y el ataque consiste en suponer una determinada longitud de clave empezando desde longitud 1 hasta la longitud que supongamos pueda ser máxima para una llave, posterior a esto, para cada una de estas elecciones de longitud se calcula la entropía del mensaje como se muestra en el algoritmo de "entropía total" de la figura 2 y al final se toma la longitud de clave cuyo valor de entropía sea menor o igual a la entropía del lenguaje. Esta será la longitud de clave del mensaje encriptado. Este método se puede justificar porque la entropía del texto claro se conserva en cada uno de los conjuntos C_j cuando se ha asumido la longitud correcta de la llave. [5]

Es importante mencionar que el algoritmo "Particion'' de la figura 2 se ocupa de dividir el mensaje originalmente cifrado con el método de Vigenère en k mensajes cifrados con el método del Cesar. La función H usada en el algoritmo ''entropia total'' es la definida en (6).

V. ATAQUE DE MÁXIMA CORRELACIÓN

Los algoritmos criptográficos se pueden clasificar en algoritmos de flujo y algoritmos en bloque [1]. Los algoritmos en flujo se basan en generar una secuencia pseudo-aleatoria con la que se pueda encriptar el mensaje. Los algoritmos en bloque en cambio utilizan alguna función que dado un bloque de símbolos del texto claro y una llave, calculan el bloque de símbolos correspondientes en el texto encriptado.

function Particion(k, $texto$)

1 **para** $i = \{0, ..., k-1\}$
2 $C_i = \emptyset$
3 $j = 0$
4 **mientras** $(jk + i) < M$
5 $C_i = C_i \cup \{\text{texto}[jk + i]\}$
6 $j = j + 1$
7 **retornar** $< C_0, C_1, ..., C_{k-1} >$

function EntropiaTotal(k, $texto$)

1 $C = \text{Particion}(k, texto)$
2 $sum = 0$
3 **para** $i = \{0, ..., k-1\}$
4 $sum = sum + H(C_i)$
5 **retornar** $\left[\frac{sum}{k}\right]$

Figura 2

Formalmente se puede decir que un algoritmo de cifrado en bloque, convierte un símbolo o bloque de símbolos de texto claro S_i en un símbolo o bloque de símbolos del texto encriptado C_i con una función f de la siguiente manera:

$$C_i = f(S_i, K) \qquad (7)$$

Donde K es la llave utilizada para el cifrado, de la misma manera, para hacer el descifrado del texto se utilizaría la siguiente ecuación:

$$S_i = f^{-1}(C_i, K) \qquad (8)$$

Como se explicó en la sección 4, el cifrado de Vigenère puede ser dividido en varios cifrados del cesar una vez se conoce la longitud de la clave. En esta sección se explica cómo romper cada uno de los cifrados del Cesar una vez realizado el procedimiento anterior. El algoritmo del Cesar, se puede ver como un cifrado en bloque cuya función f está dada por:

$$f(S_i, K) \equiv S_i + K \ (mod\ N) \qquad (9)$$

Partiendo de la suposición que conocemos las frecuencias relativas $F_1(S_i)$ con las que aparece cada símbolo S_i en el lenguaje del texto claro, el objetivo se vuelve conocer cuál es el parámetro K que maximiza la similitud de las frecuencias $F_1(S_i)$ y $F_2(f(S_i, K))$. A esta medida de similitud basada en las frecuencias relativas de aparición de cada símbolo en el lenguaje se le conoce como correlación, la correlación G para una determinada clave k está dada por:

$$G(k) = \sum_{i=0}^{N-1} F_1(S_i) * F_2(f(S_i, k)) \qquad (10)$$

Donde $F_1(X)$ es la frecuencia del símbolo X en el lenguaje en el que se encuentra el texto claro y $F_2(X)$ es la frecuencia con la que se encuentra el símbolo X en un texto cifrado con el método del Cesar (un conjunto C_j).

Dado que k puede tomar valores entre 0 y N (Donde N es la cantidad de símbolos del lenguaje), se pueden calcular $G(k)$ para todo k entre 0 y N, y tomar como la llave el k que maximice $G(k)$ [5].

VI. EJEMPLO

En esta sección se muestra un ejemplo del funcionamiento de todo lo explicado en este artículo. A continuación se presenta el texto a cifrar:

LOSALGORITMOSCRIPTOGRAFICOSUTILIZADOSDESD
ELAEPOCADELAANTIGUAROMAHASTANUESTROSDIA
SSONMETODOSQUECONVIERTENUNMENSAJEDETEXT
OPLANOENTEXTOCIFRADOELPROCESOINVERSOSECO
NOCECOMODESCIFRARYCONSISTEENLLEVARELTEXT
OCIFRADOATEXTOCLAROUSUALMENTEESTOSALGOR
ITMOSUTILIZANUNALLAVESECRETACOMOPARTEDEL
AENTRADAENLOSMETODOSQUESEVANAMOSTRAREN
ESTEARTICULOCONOCIDOSCOMOCIFRADODELLAVES
IMETRICALALLAVEDEBECOINCIDIRTANTOENELPROC
ESODECIFRADOCOMOENELDEDESCIFRADOPARAQUE
LACOMUNICACIONENTREELEMISORQUIENCIFRACOM
OELRECEPTORQUEDESCIFRASEAEXITOSA.

Al cifrar el texto anterior utilizando el método de Vigenère con una determinada clave se obtuvo el siguiente texto cifrado.

TBKIWOQFQGEWDKTWXGGOCIHWKBKCEQNWH
NVWDLGGLRDIPXQQIQWTLIPHQTMICWOOPNKB
LVWSAGJWDLKOAFGVXMVCLBKYFMECVIAMCB
GBCAEMYACXMQWBPFVCXYSVZMPHMKLWNQH
FIQGMWXTCKRKWTVXSZFGAPKQBWPWKZUQRM
FUQQZCFGPGVDQUHMRFTWMXOZRDBPFVCKVX
ZLLQOBRPBZKNOZBMAFINAMALMPAVCANDOZZ
KHUBKCEQNWHNFCYINZIIWAPKTSBNUWXWROZ
GWLPTCSVGJIOIGBTBKUPBQRWFICPAGJIASUZAV
FIEWVPAVSIELQNCNCKBFWNQFCAPGUZKKTZNV
WOMNZIIWATUGHZVUIWINZIIWLPJGQWVFKTLK
FBNFBZMPSTCJWNMUCLRUQQZCRWPGUZMPSTQ
WLPAEWNESLZXCFIDMMWIECUHFQNIEWWAWV
EZGSTREQDWTECVWVNQHFIPGUZMNFMPWXEW
TECRVMDKKTZNKMLMZWBBKI

Para atacar el método criptográfico, se utiliza primero la entropía para hallar la longitud de la clave como se explicó en la sección 4. Al correr el algoritmo de la figura 2 para diferentes valores de k, se obtienen las entropías de la tabla 4.

K	EntropiaTotal (K, texto)
1	3.1591
2	3.0878
3	3.1056
4	2.9287
5	3.0717
6	2.9467
7	2.9703
8	2.5717
9	2.9245

Tabla 4.

Se realizó el cálculo de la entropía del libro "Don quijote de la mancha" para estimar la entropía del español, obteniendo un valor de entropía cercano a 2,8. Como se puede observar en la tabla 4, el primer valor de K cuya entropía es menor o igual a la del lenguaje es K=8. Así que por ahora la hipótesis es que le longitud de la clave es 8 caracteres. Luego se deben partir el mensaje en 8 conjuntos, colocando el primer carácter en el primer conjunto (C0), el segundo en el segundo conjunto (C1),…, el noveno en el primer conjunto (C0), el décimo en el segundo y así sucesivamente. En la tabla 5 se muestra como se dividen los primeros 16 caracteres.

C0	C1	C2	C3	C4	C5	C6	C7
T	B	K	I	W	O	Q	F
Q	G	E	W	D	K	T	W

Tabla 5

Para el cálculo de máxima correlación se obtuvieron las frecuencias de aparición de cada carácter del libro del Quijote de la Mancha. Al calcular la correlación para todos los valores de k para el primer conjunto C0 se obtiene un máximo en k=8, lo que indica que el primer carácter de la clave es 'I' como se muestra en la tabla 1, al hacer este procedimiento para cada conjunto hasta llegar a C7, se obtiene que la clave de cifrado es "INSILICO".

VI. RESULTADOS

El algoritmo del cálculo de la entropía para diferentes longitudes de clave fue implementado en el lenguaje de programación Python. Por otro lado, los algoritmos de cifrado y descifrado de Vigenère, y el de máxima correlación fueron implementados en el lenguaje de programación C++. Se probaron los métodos anteriormente descritos sobre textos de diferente longitud y diferente tamaño de clave, los resultados se muestran en la tabla 2. En esta tabla se muestra la efectividad del algoritmo para mensajes de diferentes longitudes y diferentes longitudes de clave, donde la efectividad del algoritmo se mide como la cantidad de caracteres de la clave hallados exitosamente por el algoritmo sobre la cantidad total de caracteres de la clave. Se puede observar que entre más largo sea el mensaje, mayor precisión tiene el análisis estadístico, y que entre más larga sea la clave, el análisis tiene una menor precisión.

El algoritmo mostrado en este artículo tiene una complejidad algorítmica $\Theta(KN)$ mientras que por fuerza bruta la complejidad sería $\Theta(K^N)$, siendo N la cantidad de dígitos de la clave y K la cantidad de símbolos del alfabeto.

tamaño mensaje	tamaño clave	% efectividad
364	4	75%
500	4	100%
500	5	100%
500	8	100%
500	22	77,20%
1000	4	100%
1000	9	100%
1000	16	100%
1000	22	86,30%
3041	22	100%

Tabla 6.

VII. CONCLUSIONES Y RECOMENDACIONES

El uso de conceptos de la teoría de la información junto con herramientas estadísticas permite explorar nuevas formas de hacer criptoanálisis tanto en los métodos clásicos como en la criptografía moderna, además constituye una alternativa práctica de análisis y de rápida implementación.

REFERENCIAS

[1] Menezes, A.J. and Van Oorschot, P.C. and Vanstone, S.A "Handbook of applied cryptography" 1997

[2] Luciano, D. and Prichett, G. "Cryptology: From Caesar ciphers to public-key cryptosystems" The College Mathematics Journal, vol 18 pp 2-17, 1987

[3] Van Tilborg, H.C.A. "Encyclopedia of cryptography and security" pp 115, 2005

[4] Smith, L.D. "Cryptography: the science of secret writing" 1955

[5] Sabater, F.A Guia, M.D. Hernandez, E.L. Montoya V.F. Muñoz, M.J. "Tecnicas criptográficas de protección de datos" pp 12-27, 1997

[6] Knudsen, Lars R. (1998). "Block Ciphers— a survey" June 1997

[7] David, Kahn (1999) "The Codebreakers: The Story of Secret Writing" 999

[8] Sinkov, Abraham; Paul L. Irwin "*Elementary Cryptanalysis: A Mathematical Approach*" Mathematical Association of America 2000