Ataque de análise de Frequência na Cifra de Vigenere

Guilherme Mattos Camargo, 17/0104508

[1]Dep. Ciência da Computação – Universidade de Brasília (UnB)
CIC 0201 - Segurança Computacional

170104508@aluno.unb.br

1. Introdução

A cifra de Vigenere é um método de criptografia polialfabético, onde uma palavra chave é escolhida para cifrar uma mensagem. A cifra ocorre pelo deslocamento de cada caractere do texto pelo número correspondente a cada letra (a=0, b=1 ... z=25) da palavra chave, quando a palavra chave acaba, ela é repetida para o próximo caractere da mensagem. O modelo matemático para cifragem e decifragem pode ser visto abaixo:

- C_i: Caractere cifrado
- M_i: Caractere não cifrado da mensagem
- P_i: Caractere da palavra chave

$$Cifragem : Ci = Mi + Pi(mod26) \tag{1}$$

$$Decifragem : Mi = Ci - Pi(mod26) \tag{2}$$

2. Vulnerabilidade

Existem características fundamentais da cifra de Vigenere que devem ser notadas ao realizar uma análise de sua segurança. Esse método não realiza permutações de caracteres e repete periodicamente a palavra chave, com isso surgem padrões muito claros que podem ser facilmente explorados.

2.1. Análise de espaçamento entre repetições

Essa análise tem o intuito de determinar o tamanho da palavra chave. Devido a repetição dessa palavra por toda a mensagem, os padrões de frequência apresentam informações extremamente úteis. Na linguagem natural, em qualquer idioma, percebe-se a repetição de diversos grupos de letras. Na mensagem criptografada essas repetições eventualmente serão cifradas pelas mesmas letras da palavra chave, com isso ao analisar o espaçamento entre as ocorrências de um grupo de caracteres, é possível chegar a diversos múltiplos do possível tamanho N da palavra chave.

2.2. Proximidade de frequência natural de aparição das letras

Essa análise tem como objetivo efetivamente determinar qual é a palavra chave, assumindo seu tamanho como N. Inicialmente deve-se dividir a mensagem criptografada em N subgrupos, onde cada grupo contém as letras que sofreram o deslocamento por uma mesma letra da palvra chave.

- Subgrupo[0] = [M[0], M[0+N], M[0+2N], ...]
- Subgrupo[1] = [M[1], M[1+N], M[1+2N], ...]
- ...

Dessa forma, para cada subgrupo, realiza-se uma comparação de distância da frequência de aparição das letras no subgrupo com a frequência de aparição natural das letras em um deteminado idioma. Essa comparação é repetida para um deslocamento correspondente a cada letra do alfabeto, de maneira que o deslocamento que proporcionar a menor diferença da frequência natural é eleito como o caractere da palavra chave.

3. Implementação

Para realizar a cifragem, decifragem e análise da cifra de Vigenere, foi implementado um código em linguagem C que pode ser encontrado em `https://github.com/gmc-b/Cifra-de-Vigenere`, as intruções de uso e outros detalhes estão presentes no arquivo README.md.
Nessa sessão serão discutidos os parâmetros utilizados e procedimentos práticos.

3.1. Análise de espaçamento entre repetições

Para essa análise percebeu-se empiricamente que os melhores resultados são obtidos utilizando combinações de 3 letras, porém é possível alterar esse valor na função dependendo da necessidade.
É estabelecido um valor mínimo de fator para se considerar, nesse trabalho o valor escolhido foi 4, pois os fatores 2 e 3 são muito comuns e acabam ”poluindo”a estatística.
São escolhidos os dois fatores com o maior valor V para teste na próxima etapa, esse valor é dado pela fórmula:

$$Valor_{fator} = Frequencia_{fator} \cdot \sqrt{Fator} \tag{3}$$

Essa fórmula é uma maneira de valorizar fatores mais altos, porém de maneira suavizada, com a raiz. Ela se mostrou eficaz no escopo de uma palavra chave de até 25 letras.

3.2. Proximidade de frequência natural de aparição das letras

O método de avaliação de proximidade utilizado foi o chi-squared:

- Fn_i: Frequência natural de aparição de i-ésima letra do alfabeto
- Fsd_i: Frequência de aparição da i-ésima letra do alfabeto no subgrupo deslocado

$$\chi^2 = \sum_{k=0}^{25} \frac{(F_n - F_{sd})^2}{F_n} \tag{4}$$

Ao determinar o menor valor de chi-squared para cada subgrupo é montada, assim, a nova senha.
Por último, ocorre uma análise de repetição na senha encontrada. Em casos de senhas que efetivamente tenham valores abaixo do mínimo estipulado é possível por exemplo que o programa encontre uma palavra chave ”abcabc”, dessa forma ele testa a repetição e caso exista, apenas metade da palvra será disponibilizada como chave (”abc”nesse caso).

3.3. Confiança

Por se tratar de um método estatístico, a precisão da resposta está atrelada ao tamanho do espaço amostral, dessa forma, para esse algoritmo, percebe-se que é possível ter respostas com boa acurácia com 100 ou mais caracteres.

4. Conclusão

A cifra de Vigenere provou-se extremamente vulnerável a um ataque relativamente simples, mostrando que, atualmente, não deve ser utilizada como método de criptografia. A análise realizada nesse trabalho, porém, tem grande valor nos aprendizado da segurança computacional, sendo possível extrapolar os métodos aqui utilizados para problemas mais complexos.