Agrupar Textos Cifrados é Equivalente a Agrupar Textos Legíveis

William A. R. de Souza[1,2], Luís Alfredo Vidal de Carvalho[1], José Antonio Xexéo[3]

[1]COPPE/UFRJ – Universidade Federal do Rio de Janeiro

Caixa Postal 68511 – CEP 21945-970 - Rio de Janeiro – RJ – Brasil

[2]Divisão de Criptologia – Centro de Análises de Sistemas Navais

Praça Barão de Ladário s/n, Centro, CEP: 20091-000– Rio de Janeiro – RJ

[3]Seção de Engenharia de Sistemas – Instituto Militar de Engenharia

Pça. General Tibúrcio 80, Praia Vermelha, CEP 22290-270 – Rio de Janeiro – RJ

{william, alfredo}@cos.ufrj.br, xexeo@ime.eb.br

Abstract. *Several studies have been made in attempt to break confidentiality, either by obtaining the knowledge of the plaintext or the key itself working only with cryptograms. However, there is not known methods capable of breaking contemporary cryptographic algorithms, as DES and AES. Nevertheless, in order to benefit cryptanalysts, it is possible to search weakness in these algorithms. In this work we show that ciphertexts can be considered as plaintexts written in an unknown idiom and using a binary alphabet, where each idiom is determined by the cryptographic key. In the experiments with ciphertexts and plaintexts clustering it have reached success, since all ciphertexts encrypted with the same key belong to the same group, as well as, plaintexts, written in the same idiom and alphabet belong to the same group. This result exposes a cryptographic algorithms weakness, since they are designed to generate ciphertexts without any relation with the input data, such as the plaintext or the cryptographic key.*

Resumo. *Muitos estudos têm sido realizados na tentativa de comprometer o sigilo determinado por algoritmos criptográficos para obtenção do conhecimento do texto legível ou da própria chave, com o conhecimento apenas do criptograma gerado. Contudo, não são conhecidos métodos capazes de quebrar algoritmos criptográficos atuais, como o DES e o AES. Apesar disso, em benefício dos criptoanalistas, é possível procurar fraquezas nesses algoritmos. Neste trabalho mostramos que os textos cifrados podem ser considerados como textos legíveis escritos em um idioma desconhecido e utilizando um alfabeto binário, onde cada idioma é determinado pela chave criptográfica. Nos experimentos com agrupamento de textos cifrados e legíveis foi alcançado sucesso, ocorrendo o fato de textos cifrados com a mesma chave serem reunidos no mesmo grupo, assim como, textos legíveis escritos em idiomas e alfabetos iguais. Esse resultado expõe uma fraqueza dos algoritmos criptográficos, já que os mesmos são projetados para gerar textos cifrados sem qualquer relação com os dados de entrada, como a chave criptográfica.*

1. Introdução

O objetivo principal da criptografia é manter o sigilo de mensagens que trafegam em um canal de comunicações inseguro. Assim, em um canal sujeito a interceptação, usar criptografia é fundamental para a preservação do sigilo, de tal maneira que somente as partes que possuem a chave criptográfica (utilizada para cifrar e/ou decifrar as mensagens) tenham acesso ao conteúdo de uma mensagem cifrada.

Um sistema criptográfico pode ser definido como uma quíntupla (P, C, K, E, D), onde as seguintes condições são satisfeitas [Stinson 2006]:

1. P é um conjunto finito de possíveis textos legível;

2. C é um conjunto finito de possíveis textos cifrados;

3. K, o espaço de chaves, é um conjunto finito de possíveis chaves;

4. Para cada $k \in K$, existe uma regra de cifragem $e_k \in E$ e uma correspondente regra de decifragem $d_k \in D$. $e_k : P \to C$ e $d_k : C \to P$ são funções, tais que $d_k(e_k(x)) = x$ para todo texto legível $x \in P$.

O seu uso remonta a mais de 4000 anos [Kahn 1967], mas só a partir da década de 1970, o NBS[1] estabeleceu o primeiro algoritmo criptográfico padrão com base em cifras[2] simétricas de blocos, o DES[3] [NIST 1999]. Em 2001, o NIST estabeleceu o novo padrão criptográfico, também baseado em cifras simétricas de blocos, denominado AES[4] [NIST 2001], baseado no algoritmo Rijndael [Daemen e Rijmen 2002].

Para tornarem-se padrões, esses algoritmos passaram em ensaios de aleatoriedade dos textos cifrados; isto é, independentemente da chave ou do texto legível (mensagem), o texto cifrado (criptograma[5]) deve ter características bem próximas de um texto sorteado ao acaso, sem conter qualquer padrão que possa revelar informações sobre a mensagem ou sobre a chave utilizada. O AES, por exemplo, foi certificado por uma bateria de testes do NIST (2001).

Esses testes, entretanto, não excluem a possibilidade de que outro tipo de ensaio identifique, nos criptogramas, padrões relacionados à chave, à mensagem ou ao próprio algoritmo. Por exemplo, criptogramas gerados pelo algoritmo RC6 podem ser identificados com o auxílio da estatística do χ^2, no chamado "ataque de distinção" [Knudsen e Meier 2000], assim como também é possível agrupar criptogramas em função da chave de cifrar [Souza et al 2008]. Há também alguns relatos de busca por padrões em criptogramas [Souza 2007]. Isso mostra a possibilidade de se obter informações importantes a partir dos criptogramas.

A criptoanálise é o ramo da criptologia que busca obter o conhecimento do texto legível ou da própria chave, sem o conhecimento da chave criptográfica. A criptoanálise clássica era fortemente baseada nas características lingüísticas do idioma de origem do

---

[1] National Bereau of Standards (atual NIST – National Institute of Standard and Technology).
[2] Algoritmo criptográfico.
[3] Data Encryption Standard.
[4] Advanced Encryption Standard.
[5] Neste trabalho criptograma é o mesmo que texto cifrado e mensagem é o mesmo que texto legível.

texto legível [Singh 2003]. Desta maneira, a criptoanálise explorava as propriedades intrínsecas do idioma refletidas nos criptogramas. Os métodos e técnicas utilizados em lingüística são de interesse para a criptoanálise, pois embora as técnicas atuais de criptografia envolvam problemas matemáticos difíceis, a sua matéria-prima continua sendo o texto[6] [Souza 2007].

Assim, este trabalho demonstra que os textos cifrados podem ser tratados como textos legíveis escritos em um idioma desconhecido e utilizando um alfabeto binário, onde cada idioma é determinado pela chave criptográfica utilizada no processo de cifrar. Desta forma, cada chave determina um conjunto léxico para este idioma desconhecido.

Foram realizados experimentos com agrupamento de criptogramas e mensagens, sem conhecimento das mensagens ou das chaves utilizadas, onde o procedimento foi capaz de separar tanto as mensagens quanto os criptogramas em diversos conjuntos, de tal forma que os criptogramas gerados pela mesma chave – e somente eles – ficaram agrupados no mesmo conjunto, da mesma forma que os textos legíveis de um mesmo idioma e alfabeto.

Com a finalidade de mostrar a influência dos elementos léxicos [Jurafsky e Martin 2009] no processo de agrupamento, foram selecionados textos legíveis escritos em idiomas e a alfabetos diferentes, além dos conjuntos de textos cifrados. Desta forma, nos experimentos foram utilizados dois conjuntos de 30 criptogramas gerados pelo algoritmo AES, cada um cifrado com chave diferente e aleatória (chave de 128), e dois conjuntos de 30 criptogramas gerados pelo algoritmo DES, cada um cifrado com chave diferente e aleatória (chave de 64)[7] e 30 textos legíveis de cada um dos idiomas a seguir: alemão, dinamarquês, holandês, espanhol, francês, grego, hebreu e português.

O trabalho está organizado da seguinte maneira: na seção 2 é apresentada a descrição do procedimento. Os experimentos e os resultados obtidos estão na seção 3. A conclusão constitui a seção 4.

2. Descrição do procedimento

As cifras de blocos transformam mensagens em criptogramas operando sobre blocos de bits ou bytes. O tamanho dos blocos depende do algoritmo e, muitas vezes, do tamanho da chave utilizada. O AES, por exemplo, utiliza blocos de 128, 192 ou 256 bits, dependendo do tamanho da chave[8, 9].

O procedimento considera os blocos dos criptogramas como palavras de um texto de um idioma desconhecido[10] e agrupa os criptogramas e as mensagens com base na similaridade existente entre eles, similaridade essa medida com base na freqüência com que os blocos e as palavras ocorrem em cada um dos criptogramas e das mensagens.

---

[6] Outros tipos de dados podem ser cifrados, como: sons, imagens e vídeos.

[7] Esses algoritmos foram escolhidos por serem padrões certificados.

[8] A descrição do AES só contempla o bloco de 128 bits. Porém, o algoritmo Rijndael pode tratar blocos de 128, 192, e 256 bits [NIST 2001].

[9] Neste trabalho, os tamanhos de chaves são iguais aos tamanhos de blocos, respectivamente.

[10] A chave determina um conjunto léxico para o idioma desconhecido em questão. Logo, os blocos dos criptogramas constituem esse conjunto.

Assim, os documentos[11] são representados de maneira que seja possível determinar similaridades entre eles, dois a dois [Rasmussen 1992], e construir uma matriz de similaridades. A partir dessa matriz é realizado o agrupamento pelo método da ligação simples. Por fim, a qualidade da categorização é avaliada por duas medidas: revocação (recall) e precisão (precision).

2.1. Medida de similaridade

Para representar os documentos são utilizados vetores de *dimensão-n*, sendo *n* o número de termos[12] do conjunto de documentos, sem repetição. Assim, sejam os vetores $d_i = (d_{1,i}, d_{2,i}, ..., d_{n,i})$ e $d_j = (d_{1,j}, d_{2,j}, ..., d_{n,j})$, dois documentos para os quais se deseja obter a similaridade. O valor relacionado à $d_{k,i}$, onde *k* representa o *k-ésimo* termo de d_i, é a freqüência do termo *k* em d_i. O valor relacionado à $d_{k,j}$, onde *k* representa o *k-ésimo* termo de d_j, é a freqüência do termo *k* em d_j.

A similaridade entre dois documentos d_i e d_j foi associada ao ângulo do co-seno [Harman 1992] e calculada pela fórmula (1). Quanto maior o valor de S, maior a similaridade entre os documentos. Assim, constrói-se uma matriz de similaridades armazenando, em suas células, os valores de similaridade dos pares de documentos na coleção.

$$S_{Co-seno}(d_i, d_j) = \frac{\sum_{k=1}^{n} (d_{i,k} \times d_{j,k})}{\sqrt{\sum_{k=1}^{n} (d_{i,k})^2 \times \sum_{k=1}^{n} (d_{j,k})^2}} \quad (1)$$

2.2. Agrupamento

O agrupamento utiliza a técnica hierárquica aglomerativa da ligação simples [Rasmussen 1992] formando grupos dispersos [Jain 1999]. Assim, no final do procedimento, *n* documentos são agrupados em *m* grupos, onde o valor de *m* é normalmente desconhecido.

No início do processo, cada documento pertencerá a um único grupo. A seguir, identifica-se, na matriz de similaridades, o par de documentos com o maior valor de similaridade para formar o primeiro grupo. Atribui-se ao grupo um valor de "similaridade de grupo" igual ao maior valor de similaridade existente entre os pares de documentos pertencentes ao grupo. A matriz de similaridade é atualizada pela substituição da similaridade do par pela similaridade do grupo. Esse procedimento é repetido até que um critério de parada seja alcançado.

A estrutura final do agrupamento, representando a ordem em que as inclusões e junções dos grupos ocorrem é representada por um dendrograma [Rasmussen 1992] (figura 2). Observando essa figura, pode-se notar que os grupos foram formados a partir

[11] O termo documento é utilizado para referenciar genericamente uma mensagem ou um criptograma
[12] Um item léxico que ocorre em uma coleção [Jurafsky 2009], neste trabalho: um bloco ou uma palavra.

de um determinado valor de similaridade, o qual pode ser utilizado como o critério de parada citado anteriormente.

Neste trabalho, utiliza-se nos experimentos um valor de similaridade próximo de zero como critério de parada, pois é pouco provável a repetição de termos ao longo dos documentos, o que torna a ligação simples adequada, uma vez que a dispersão gerada pelo método permite que dois documentos quaisquer que estejam em um grupo, possuam valor de similaridade mais baixo que a similaridade do próprio grupo.

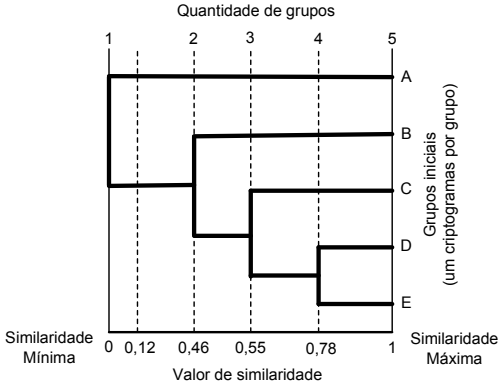

Figura 2. Dendrograma [Souza, 2008]

2.3. Avaliação dos grupos

Na avaliação da qualidade do agrupamento foram utilizadas as medidas revocação e precisão [Yates 1999] e [Fung 2003] (Figura 3). Os valores de revocação e precisão são definidos como segue. Suponha que K seja o conjunto formado pelos documentos pertencentes a uma determinada classe Δ. Seja G o agrupamento construído pelo procedimento e que supostamente contém documentos pertencentes a uma determinada classe Δ. Seja $|k|$ o número de elementos no conjunto K, $|g|$ o número de elementos de G e n o número de elementos do conjunto K presentes no grupo G. Então, os valores de Revocação e Precisão são obtidos pelas Fórmulas 2 e 3. O valor máximo para a Revocação e Precisão é 1.

Revocação, portanto, indica a capacidade do método de recuperar todos os documentos relevantes. Precisão, por sua vez, indica a capacidade do método de recuperar apenas documentos relevantes.

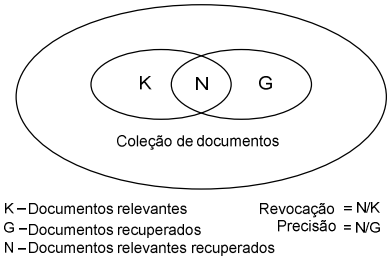

Figura 3. Revocação e Precisão

$$revocação = \frac{n}{|k|} \tag{2}$$

$$precisão = \frac{n}{|g|} \tag{3}$$

3. Experimentos, avaliações e resultados[13]

O objetivo dos experimentos é verificar se aplicando o procedimento explicado na seção 2 a uma coleção de documentos podemos obter grupos, de tal forma que, somente criptogramas cifrados com uma mesma chave pertençam ao mesmo grupo e somente mensagens escritas em um mesmo idioma pertençam ao mesmo grupo. As mensagens foram obtidas em [Bible 2009] e os resultados gerados pela ferramenta WARSText.

3.1. Experimento 1

O objetivo do experimento 1 é verificar o agrupamento somente das mensagens. Foram utilizados 30 textos de cada um dos idiomas a seguir: alemão, dinamarquês, holandês, espanhol, francês, grego, hebreu e português. Para identificar o valor máximo de precisão e revocação, foram testados três valores de corte (tabela 1).

Tabela 1. Resultados de *precisão* e *revocação* para o experimento 1

Oito idiomas diferentes, sendo dois com alfabetos distinto dos demais.		
Corte	Precisão	Revocação
0,001	0,375	1
0,455	1	1
0,550	1	1

Observando a tabela 1, pode-se concluir que o agrupamento ocorreu com sucesso, obtendo o valor máximo de precisão e revocação a partir da similaridade de corte 0,455. Os idiomas utilizados no experimento indicam a existência de nove grupos naturais, cada grupo representado por um idioma. Uma inspeção no conteúdo dos textos pode sugerir a existência de quatro grupos representados por um valor de similaridade qualquer no [0,1], com a configuração abaixo:

1. Grupo 1: alemão, dinamarquês, holandês;
2. Grupo 2: espanhol, francês e português;
3. Grupo 3: grego; e
4. Grupo 4: hebreu.

A formação acima se deve a maior ou menor interseção de termos entre os textos, o que indica que à medida que o valor de corte tende a zero a precisão é reduzida

[13] Os criptogramas, as mensagens, os arquivos contendo os grupos formados e as ferramentas estão disponíveis em http://www.portalcomputacao.com.br/linguistica.html.

uma vez que uma menor quantidade de interseções e requerida para gerar a pertinência a um grupo, degenerando a um grupo quando do valor zero.

Por outro lado, quando o valor de corte tende a um, a precisão aumenta, já que uma interseção cada vez maior de termos é exigida para a pertinência a um grupo, levando ao caso trivial onde cada grupo possui apenas um texto.

Observando-se os grupos formados com o valor de corte 0,001, nota-se que embora o processo tenha gerado grupos imprecisos, os textos em grego e em hebreu não se misturaram em grupos de outros idiomas, o que reduz o processo de agrupamento aqui descrito a um fato estatístico [Rasmussen 1992] [Jain 1999].

3.2. Experimento 2

O objetivo do experimento 2 é verificar o agrupamento somente dos criptogramas. Foram utilizados dois conjuntos de 30 criptogramas cada, gerados pelo algoritmo AES, cada conjunto cifrado com chave diferente e aleatória (chave de 128), e dois conjuntos de 30 criptogramas cada, gerados pelo algoritmo DES, cada conjunto cifrado com chave diferente e aleatória (chave de 64).

Tabela 2. Resultados de *precisão* e *revocação* para o experimento 2

Dois conjuntos de criptogramas cifrados com o algoritmo AES e dois cifrados com o DES.		
Corte	Precisão	Revocação
0,001	1	1
0,455	1	0,333
0,550	1	0,333

Observando a tabela 2, pode-se concluir que o agrupamento ocorreu com sucesso, obtendo o valor máximo de precisão e revocação com a similaridade de corte 0,001.

Considerando o princípio de que uma chave determina um idioma, o qual possui o seu próprio conjunto léxico, pode-se deduzir a existência de quatro grupos naturais, cada grupo representado por uma chave criptográfica utilizada no processo de cifrar, duas para o algoritmo AES e duas para o algoritmo DES.

O experimento demonstra que os quatros grupos são formados a partir do valor de corte 0,001. Considerado o fato de que poucos termos se repetem ao longo dos criptogramas, é esperado que mesmo criptogramas com baixa interseção de termos fiquem no mesmo grupo. Este fato, somado ao fator de dispersão explicado na seção 2.2 justificam o sucesso do agrupamento com um baixo valor de corte.

No caso dos criptogramas, a baixa interseção entre estes leva o resultado do processo rapidamente ao caso trivial onde cada grupo possui apenas um criptograma, quando o valor de corte tende a um.

3.3. Experimento 3

O objetivo do experimento 3 é verificar o agrupamento a partir de toda coleção, isto é, os dados do experimento 1 e do experimento 2.

Tabela 3. Resultados de *precisão* e *revocação* para o experimento 3

Dados do experimento 1 e 2.		
Corte	Precisão	Revocação
0,001	0,583	1
0,455	1	1
0,550	1	1

Considerando a hipótese de que os criptogramas podem ser considerados textos de um idioma desconhecido e determinado pela chave criptográfica e observando-se os resultados dos experimentos um e dois, tem-se que juntar os dados destes experimentos equivale a ter-se 13 grupos naturais, onde cada grupo possui documentos escritos em idiomas diferentes, e onde alguns desses idiomas utilizam alfabetos diferentes (binário, grego e hebraico).

Assim, reduz-se o problema do agrupamento deste conjunto de dados ao número de interseções de termos ao longo dos documentos.

O agrupamento ocorreu com sucesso (tabela 3), obtendo o valor máximo de precisão e revocação a partir da similaridade de corte 0,455.

3.4. Experimento 4

Um último experimento foi necessário para verificar se somente com mensagens escritas com alfabetos diferentes, ou seja, sem repetir idiomas que utilizam o mesmo alfabeto, e mais os conjuntos de criptogramas, o agrupamento poderia ser realizado com sucesso. Foram utilizados os dados do experimento 2 e 30 mensagens de cada um dos idiomas a seguir: português, grego e hebreu.

Tabela 4. Resultados de precisão e revocação para o experimento 4

Dados do experimento 2, mais os idiomas português, grego, hebreu.		
Corte	Precisão	Revocação
0,001	1	1
0,455	1	1
0,550	1	1

Na tabela 4, observa-se que o agrupamento ocorreu com sucesso, com precisão e revocação máxima para todas as similaridades de corte. Mais uma vez, reduz-se o problema do agrupamento do conjunto de dados ao número de interseções de termos ao longo dos documentos. Neste caso, a interseção entre os sete grupos naturais, é zero, dado que os termos estão todos escritos em alfabetos diferentes.

4. Conclusões

O trabalho propõe que os textos cifrados podem ser tratados como textos legíveis escritos em um idioma desconhecido e utilizando um alfabeto binário, onde cada idioma é determinado pela chave criptográfica. E demonstra que tal proposta é válida ao agrupar criptogramas, de tal forma que criptogramas originados da mesma chave agrupam-se num mesmo conjunto e cada conjunto só contém criptogramas gerados com a mesma chave, da mesma forma que os textos legíveis, escritos em idiomas e alfabetos iguais, também se reúnem em um mesmo grupo. Pode-se comprovar que o agrupamento depende apenas do conjunto léxico, sem levar em conta as características sintáticas e semânticas dos documentos, o idioma em que estão escritos ou se eles são legíveis ou não.

Alguns trabalhos têm buscado identificar padrões em criptogramas com técnicas de lingüística computacional, como pode ser visto em [Souza et al 2008] e em [Souza 2007]. Técnicas estatísticas, como o uso da estatística do χ^2, podem ser utilizadas para obter resultados semelhantes. Este resultado motiva o estudo de técnicas complementares para a certificação de algoritmos criptográficos.

Referências Bibliográficas

Bible (2009), "The unbound bible", Disponível: http://unbound.biola.edu [capturado 13 abr. 2009].

Daemen, J. and Rijmen V. (2002), The design of Rijndael: AES – the Advanced Encryption Standard. Springer.

Fung, B. C. M., Wang, K. e Ester M. (2003), Hierarchical document clustering using frequent itemsets. *Proceedings of the SIAM International Conference on Data Mining*, San Francisco.

Harman, D. (1992), "Ranking algorithms". In Information retrieval: data structures and algorithms, Edited by William Frakes and Ricardo Yates, Prentice Hall, p. 363–392.

Jain, A. K, Murty, M. N e Flynn, P. J (1999). *Data Clustering: A Review*. ACM Computing Surveys, Vol. 31, No 3, Setember 1999. p. 264-323.

Jurafsky, D. and Martin, J. H. (2009), Speech and language processing: An introduction to natural language processing, computacional linguistics, and speech recognition. 2th ed. Pearson.

Kahn, D. (1967), The codebreakers: the story of secret writing. Macmillan Publish1ing.

Knudsen, L.R. and Meier, W. (2000). *Correlations in RC6 with a Reduced Number of Rounds. Proceedings of the 7th International Workshop on Fast Software Encryption.*

NIST (1999). Federal Information Processing Standard, publication 46-3 (FIPS 46-3): Data Encryption Standard (DES). Washington D.C.

NIST (2001). Federal Information Processing Standard, publication 197 (FIPS 197): Announcing the advanced encryption standard (AES). Washington D.C.

Rasmussen, E. (1992), “Clustering algorithms”. In Information retrieval: data structures and algorithms, Edited by William Frakes and Ricardo Yates, Prentice Hall, p. 419-442.

Souza, W. A. R. de (2007). Identificação de padrões em criptogramas usando técnicas de classificação de textos. Dissertação de Mestrado – Instituto Militar de Engenharia. Disponível em: http://www.cos.ufrj.br/~william/teseIME.pdf

Singh, S. (2003), O livro dos códigos. Record.

Souza, W. A. R. de; Xexéo, J. A. and Oliveira, C.M.G.M. (2008). *Método de Agrupamento de Criptogramas em Função das Chaves de Cifrar. Anais do IV Workshop em Algoritmos e Aplicações de Datamining.*

Stinson, D. R. (2006), Cryptography: theory and practice. 3th ed. CRC.

Yates, R.B. e Neto, B. R. (1999), Modern information retrieval. Addison Wesley.