MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
SECRETARIA DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
CURSO DE MESTRADO EM SISTEMAS E COMPUTAÇÃO

CARLOS ANDRÉ BATISTA DE CARVALHO

O USO DE TÉCNICAS DE RECUPERAÇÃO DE INFORMAÇÕES EM CRIPTOANÁLISE

Rio de Janeiro
2006

INSTITUTO MILITAR DE ENGENHARIA

CARLOS ANDRÉ BATISTA DE CARVALHO

O USO DE TÉCNICAS DE RECUPERAÇÃO DE INFORMAÇÕES EM CRIPTOANÁLISE

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: José Antônio Moreira Xexéo, D.Sc.
Co-orientador: Cláudia Maria Garcia de Oliveira, Ph.D.

Rio de Janeiro
2006

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80-Praia Vermelha
Rio de Janeiro-RJ CEP 22290-270

C331u Carvalho, Carlos André Batista de
O Uso de Técnicas de Recuperação de Informações em Criptoanálise, Carlos André Batista de Carvalho. – Rio de Janeiro: Instituto Militar de Engenharia, 2006.
78 p.:il., tab.

Dissertação: (mestrado) – Instituto Militar de Engenharia, Rio de Janeiro, 2006.

1. Criptoanálise. 2. Aplicações de Recuperação de Informações. I. Instituto Militar de Engenharia. II. Título.

CDD 652.8

INSTITUTO MILITAR DE ENGENHARIA
CARLOS ANDRÉ BATISTA DE CARVALHO
O USO DE TÉCNICAS DE RECUPERAÇÃO DE INFORMAÇÕES EM CRIPTOANÁLISE

Dissertação de Mestrado apresentada ao Curso de Mestrado em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para obtenção do título de Mestre em Sistemas e Computação.

Orientador: José Antônio Moreira Xexéo, D.Sc.

Co-orientador: Cláudia Maria Garcia de Oliveira, Ph.D.

Aprovada em 22 de fevereiro de 2006 pela seguinte Banca Examinadora:

José Antônio Moreira Xexéo, D.Sc. do IME - Presidente

Cláudia Maria Garcia de Oliveira, Ph.D. do IME

Almir Paz de Lima, M.Sc. do IME

Luís Alfredo Vidal de Carvalho, D.Sc. da COPPE/UFRJ

Geraldo B. Xexéo, D.Sc. da COPPE/UFRJ

Rio de Janeiro
2006

A memória de minha avó Duzinha, que veio a falecer no meio dessa minha jornada. Um verdadeiro exemplo de pessoa e de mãe que sempre lutou para fornecer o melhor para toda sua família. Essa vitória também é sua. Obrigado!

AGRADECIMENTOS

Agradeço a todas as pessoas que me incentivaram e apoiaram, possibilitando meu sucesso no mestrado, em especial:

Ao meu orientador Xexéo, pela confiança em mim depositada, pela competência e empenho no desenvolvimento deste trabalho.

A Cláudia Oliveira, por sua co-orientação. Ao professor Paz de Lima, que muito colaborou para meu aprendizado.

Ao IME e a CAPES pela oportunidade e financiamento deste trabalho.

Aos meus pais Dimas e Isabel por minha educação, formação e apoio que tem alicerçado todas minhas vitórias. Aos meus irmãos Diminhas e Moara, pela amizade e amor incondicional.

A meus amigos e parentes, que mesmo na distância estiveram sempre presentes me dando força para suportar a saudade e motivação para seguir em frente.

Aos meus amigos de apartamento, Fabrício, Eduardo, Marcelo, Handrick, Arthur e Diogo, pelo companheirismo ao longo desses dois anos.

Aos meus colegas de mestrado, especialmente ao Vitor, Alexandre, Fábio, Roberto, William e Draeger, com quem caminhei junto para transpor os diversos obstáculos durante o curso.

Ao meu tio Afonso, que me acolheu nessa reta final.

A todos os mestres e funcionários do Departamento de Engenharia de Sistemas do IME.

E principalmente a Deus.

"Uma mente que se abre a uma nova idéia, jamais volta a seu tamanho original."

ALBERT EINSTEIN

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

ABREVIATURAS

AES	-	*Advanced Encryption Standard*
CBC	-	*Cipher Block Chaining*
DES	-	*Data Encryption Standard*
ECB	-	*Electronic Codebook*
k-NN	-	*k-Nearest Neighbor*
IC	-	Índice de Coincidência
IK	-	Índice Kasiski
NBS	-	*National Bureau of Standards*
NIST	-	*National Institute of Standards and Technology*
RI	-	Recuperação de Informação
SVM	-	*Support Vector Machine*

RESUMO

Este trabalho apresenta uma aplicação inovadora de técnicas de recuperação de informações em criptoanálise. Utilizando técnicas de agrupamento foi desenvolvido um novo procedimento para a determinação do período da chave no processo de criptoanálise polialfabética. Esse conhecimento foi, então, utilizado para concretizar o objetivo principal da dissertação: provar, experimentalmente, uma fraqueza nas cifras de bloco.

Os padrões lingüísticos das mensagens são, de certo modo, propagados para os criptogramas. Entretanto, esses padrões são modificados em função da chave, que pode ser considerada como uma propriedade lingüística que determina o vocabulário da nova linguagem.

O grande diferencial deste estudo é proporcionado pelo uso de técnicas de agrupamento, que permitem a identificação desses padrões por meio da separação de criptogramas de acordo com a chave usada na cifragem.

No procedimento proposto para cifras polialfabéticas, o método Kasiski é integrado às técnicas de agrupamento, gerando resultados mais eficientes que os métodos tradicionais.

O princípio fundamental das cifras de bloco é a obtenção de criptogramas com uma distribuição tão aleatórias de símbolos que não se consiga identificar uma correlação entre os dados de entrada e de saída.

Entretanto, em cifras operadas no modo ECB, a realização com sucesso do agrupamento prova a existência dessa correlação. Assim, torna-se altamente não-recomendável o uso desse modo de operação.

ABSTRACT

This work presents an innovative application of information retrieval techniques in cryptanalysis. Using clustering techniques a new procedure for key length determination in the process of cryptanalysis of polyalphabetic ciphers was developed. This knowledge was, then, used to make possible the main objective of the thesis: to prove, experimentally, a weakness in the block ciphers.

The linguistic patterns of the messages are, in certain way, propagated for the cryptograms. However, these patterns are modified in function of the key, that can be considered as a linguistic property that determines the vocabulary of the new language.

The great differential of this study is proportionate for the use of clustering techniques, that in accordance with the key allow to the identification of these patterns through the separation of the cryptograms according to the cipher key.

In the procedure proposed for polyalphabetic ciphers, the Kasiski method is integrated to the clustering techniques, generating resulted more efficient than the traditional methods.

The basic principle of the block ciphers is the attainment of cryptograms with a distribution so random of symbols that it can't identify a correlation between the input and output data.

However, in ciphers operating ECB mode, the accomplishment successfully of the clustering proves the existence of this correlation. Thus, the use in this mode of operation becomes highly not-recommendable.

1 INTRODUÇÃO

O termo criptografia tem origem grega, e significa escrita oculta ou secreta. Como disciplina científica é a investigação de métodos e técnicas que podem ser usadas para esconder o conteúdo de uma mensagem (texto em claro), produzindo uma representação da mensagem (criptograma) que é incompreensível para um leitor desautorizado (KAHN, 1967).

A criptografia envolve a conversão de um texto em claro em criptograma, por um processo chamado cifragem. O criptograma é convertido de volta no texto original pela decifragem. A cifragem compreende um algoritmo, que recebe ainda como entrada uma chave. Documentos cifrados com um mesmo algoritmo, mas com chaves distintas são, em grande parte, diferentes. A criptoanálise está relacionada aos métodos usados para determinar a chave usada, ou recuperar o texto original sem o conhecimento da mesma.

No contexto real, cifras são projetadas para satisfazer necessidades particulares: podem ser para uso manual ou computacional; podem ser implementadas em algum tipo específico de hardware; alguém pode esperar que caracteres sejam misturados na transmissão; oponentes potenciais e suas capacidades criptográficas têm que ser considerados.

1.1 MOTIVAÇÃO

Há muito se utilizam métodos lingüísticos para realizar a árdua tarefa da criptoanálise. Até meados do século XX, a criptografia esteve calcada em técnicas baseadas em características relativas à língua geradora. Isso garantia grande aplicação de procedimentos estatísticos da linguagem, o que exigia dos criptoanalistas conhecimento sobre esse ramo (SINGH, 2001).

Com o advento dos computadores, processos de cifragem como as substituições mono e polialfabéticas, isoladamente utilizados, foram sendo gradativamente abandonados em favor dos poderosos algoritmos contemporâneos de criptografia. Com isso, a adoção de critérios meramente lingüísticos também foi renegada. A aplicação de métodos baseados em freqüências de letras, dentre outros, tornou-se completamente ineficiente em face aos algoritmos que têm por princípio fundamental uma forte aleatorização, gerando criptogramas com distribuição praticamente uniforme de símbolos.

Contudo, os computadores também trouxeram uma grande evolução às técnicas de lingüística, ampliando o escopo de sua aplicação. Atualmente, procedimentos lingüísticos têm sido utilizados em inteligência artificial, compiladores e recuperação de informações textuais (MITKOV, 2005).

Entretanto, esses novos procedimentos de lingüística não vêm sendo utilizados no contexto da criptografia. Na literatura, não se tem encontrado pesquisas aliando lingüística computacional à criptoanálise. Assim, o uso de técnicas de recuperação de informações (RI) apresenta-se, nessa dissertação, como uma solução inovadora para a localização de padrões lingüísticos em criptogramas.

Os sistemas de recuperação de informações (YATES, 1999) têm dependido fortemente do modelo "saco de palavras" para a representação de documentos, negligenciando qualquer conhecimento lingüístico da linguagem em questão, mas são altamente influenciados por características estatísticas da linguagem. Muitas tarefas de RI são resolvidas supondo que um texto A, mesmo desconhecido, é mais semelhante ao texto B do que ao texto C.

A perspectiva de que os padrões lingüísticos das mensagens são, de certo modo, propagados para os criptogramas justifica a aplicação de técnicas de RI em criptoanálise. Esses padrões não são, entretanto, percebidos em virtude da chave que pode ser considerada como uma propriedade lingüística que determina o vocabulário da nova "linguagem". Assim, é necessário que os criptogramas sejam vistos como documentos comuns escritos numa língua desconhecida.

1.2 CARACTERIZAÇÃO DO PROBLEMA

A identificação de padrões em criptogramas permite a descoberta de fraquezas de sistemas criptográficos. Os sistemas clássicos de criptografia, apesar de serem pouco utilizados, ainda são alvos de pesquisas que visam a projetar métodos mais eficientes de criptoanálise.

Na criptografia clássica, pode-se destacar a criptografia de substituição mono e polialfabética. O sistema monoalfabético consiste na substituição do alfabeto original por um alfabeto cifrado; assim, na cifragem, cada símbolo do alfabeto original é substituído por um símbolo correspondente no alfabeto cifrado. A chave é definida pelo alfabeto cifrado. Existem, atualmente, métodos bastante eficientes para a criptoanálise monoalfabética, como os baseados em algoritmos genéticos (SPILLMAN, 1993).

No sistema polialfabético não existe apenas um único alfabeto cifrado. Vários alfabetos são utilizados, alternadamente, para a cifragem de um texto em claro. A chave é composta

pelos alfabetos, e pela seqüência em que são aplicados. A quantidade de alfabetos utilizada é denominada período ou tamanho da chave. Uma vez conhecido o tamanho da chave, um criptograma pode ser decifrado quebrando-o em várias partes (uma parte por alfabeto) e resolvendo cada uma das partes por criptoanálise monoalfabética.

Neste contexto, alguns procedimentos foram desenvolvidos com a finalidade de determinar o período da chave de um dado criptograma, como o Kasiski e o Índice de Coincidência (IC) (DENNING, 1982). Entretanto, estes procedimentos tornam-se mais ineficientes à medida que o período da chave aumenta em relação ao tamanho do texto. Esse problema motivou o desenvolvimento de uma novo procedimento e serviu de base para um estudo em sistemas contemporâneos de criptografia.

É de importância tecnológica fundamental o estudo em cifras modernas, em especial as cifras de bloco. Uma cifra é considerada confiável desde que atenda a um conjunto de requisitos definidos pela comunidade. Entretanto, não se tem conhecimento de nenhuma forma de avaliação que garanta a segurança de algoritmos criptográficos, os testes existentes são basicamente de reprovação, como os propostos pelo NIST (*National Institute of Standards and Technology*) (RUKHIN, 1999).

O princípio fundamental das cifras modernas é a provável eliminação da correlação entre os dados de entrada (chave e texto em claro) com os de saída (criptograma). Uma cifra de bloco é caracterizada pela cifragem de um determinado número de bits de texto em claro (um bloco) em um bloco de criptograma do mesmo tamanho.

Uma cifra de bloco pode ser operada de vários modos, sendo o modo ECB (*electronic codebook*) o mais simples. Neste, uma mensagem é quebrada em vários blocos e cada bloco é cifrado independentemente. Se um mesmo bloco é cifrado duas vezes, com a mesma chave, os dois blocos resultantes serão iguais.

A repetição de blocos permite, por meio de técnicas de agrupamento, a identificação de criptogramas cifrados com uma mesma chave. Assim, provocando um forte questionamento sobre a aleatoriedade da saída de uma cifra de bloco

1.3 ORGANIZAÇÃO DA DISSERTAÇÃO

No capítulo 2, são expostos os conceitos de criptografia essenciais para este trabalho, como as cifras polialfabéticas e de blocos, além das suas formas de criptoanálise.

O capítulo 3 descreve as técnicas de recuperação de informações utilizadas no contexto em questão.

A aplicação das técnicas de RI em criptoanálise polialfabética, e seus experimentos são descritos no capítulo 4. O uso dessas técnicas e das adaptações correspondentes em cifras de blocos, por sua vez, no capítulo 5.

Finalmente, no capítulo 6, são apresentadas conclusões da dissertação e perspectivas de sua continuação em trabalhos futuros.

2 CRIPTOGRAFIA

O uso de um sistema criptográfico (cifra) permite que duas pessoas, sejam elas Alice e Bob[1], comuniquem-se por um canal inseguro de maneira que um oponente, Eva, seja incapaz de compreender a comunicação. Quando Alice quer enviar uma mensagem para Bob, ela converte a mensagem original em um criptograma, pelo processo chamado cifragem. Bob, ao receber a mensagem, decifra o criptograma, revelando o texto em claro. Eva, mesmo podendo interceptar a mensagem, não deve ser capaz de descobrir o conteúdo da mesma (GEBBIE, 2002).

O processo de cifragem depende de uma chave, que pode ou não ser de conhecimento público, dependendo da natureza do algoritmo em questão. Entretanto na decifragem, a chave (que pode ser a mesma da cifragem) só deve ser conhecida pela parte autorizada a tomar conhecimento da informação cifrada (LAMBERT, 2004). A FIG. 2.1 apresenta um modelo do processo de comunicação. Obviamente que, se a chave para cifragem for igual a da decifragem (ou facilmente obtida da mesma), ela deve ser enviada para o destinatário por um canal seguro de comunicação.

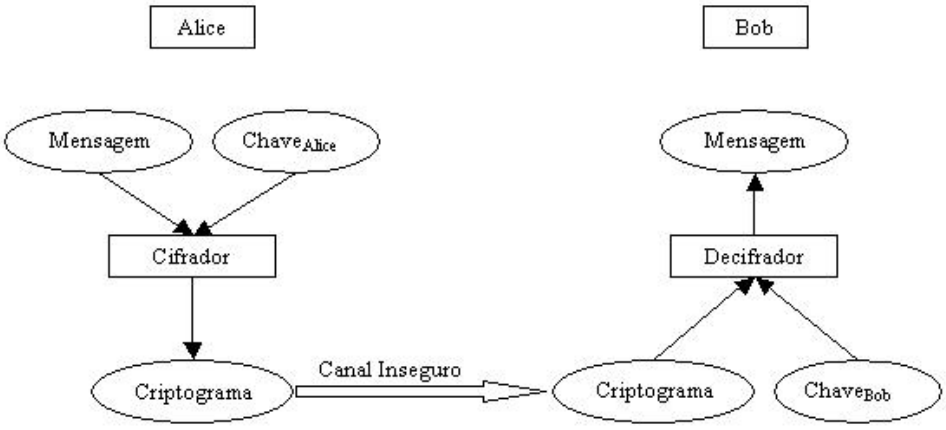

FIG. 2.1: Processo de comunicação

Matematicamente, uma cifra é uma quíntupla (ρ, γ, κ, ϵ, δ) em que as seguintes condições são satisfeitas (GEBBIE, 2002) (LINK, 2003):

[1]Nomes comumente usados na literatura.

- ρ é um conjunto finito dos possíveis textos em claro;
- γ é um conjunto finito dos possíveis criptogramas;
- κ, o espaço de chaves, é um conjunto finito das possíveis chaves;
- para cada $k \in \kappa$, existe uma regra de cifragem $e_k \in \epsilon$ e uma regra correspondente de decifragem $d_k \in \delta$. $e_k : \rho \rightarrow \gamma$ e $d_k : \gamma \rightarrow \rho$ são funções em que $d_k(e_k(x)) = x$ para todo texto em claro $x \in \rho$.

É importante salientar que a função de cifragem ek deve ser injetiva, ou seja, para todas as mensagens $x1, x2 \in \rho$, que satisfaçam $e_k(x1) = e_k(x2)$, tem-se $x1 = x2$. Caso contrário, a decifragem não poderia ser realizada de maneira não ambígua (LINK, 2003).

Um sistema criptográfico pode ser classificado quanto ao segredo da chave. Quando a chave utilizada para decifrar é igual à usada na cifragem ou facilmente obtida a partir dela, o sistema é **simétrico** (ou de chave secreta). Neste caso, a chave deve ser mantida em sigilo e trocada por um canal seguro de comunicação. Por outro lado, os algoritmos em que a chave de decifragem é dificilmente obtida a partir da chave de cifragem, são chamados **assimétricos** (ou de chave pública). Nesses sistemas, a chave utilizada para cifrar pode ser de conhecimento público (chave pública), pois a segurança esta associada ao segredo da chave usada para decifrar (chave privada) (LAMBERT, 2004).

Os algoritmos podem ser divididos ainda quanto à forma como agem sobre uma mensagem. As cifras de fluxo cifram a mensagem símbolo a símbolo, de acordo com o fluxo de entrada dos dados. Já as cifras de blocos dividem o texto em claro em blocos, cifrando cada um em separado e concatenando o resultado para compor o criptograma (LAMBERT, 2004).

Em conjunto com a criptografia surgiu a criptoanálise, que busca vulnerabilidades em um sistema criptográfico objetivando a recuperação do texto em claro sem o conhecimento da chave ou a recuperação da chave em si. Criptologia refere-se ao campo de estudo que abrange a criptografia e a criptoanálise (SCHNEIER, 1996). Atualmente, além do sigilo, a criptografia estuda técnicas para garantir a integridade e a autenticidade de uma mensagem (MENEZES, 1996).

2.1 BREVE HISTÓRICO

A arte da escrita secreta é utilizada desde a antiguidade, principalmente por entidades militares, igreja e governo. A possibilidade de interceptação de uma mensagem tem motivado a evolução da criptografia (SINGH, 2001).

Uma das primeiras técnicas utilizadas foi a de **esteganografia** (do grego, grafia coberta), que consiste em ocultar de alguma forma a mensagem enviada. No ano 480 a.C., o grego Demarato escreveu uma mensagem, alertando a Grécia sobre um ataque dos persas, em tabuletas de madeiras e as cobriu com cera, para que as mesmas pudessem chegar em segurança, sendo revelado seu conteúdo quando as tabuletas fossem raspadas. Uma das formas mais conhecidas de esteganografia é a tinta invisível.

Ainda na antiguidade desenvolveu-se a criptografia propriamente dita que, em contraste com a esteganografia, ocultava o conteúdo da mensagem, ou seja, podia-se ver a mensagem, mas não compreendê-la. Existem dois meios de se codificar uma mensagem: por **transposição** ou **substituição**. Uma transposição consiste em reorganizar as letras de uma mensagem, formando um anagrama. Ou seja, as posições das letras no texto são simplesmente alteradas. No séc. V a.C., a Grécia utilizou um aparelho, chamado *citale*, para enviar mensagens criptografadas. O mesmo consistia em um bastão de madeira, que, quando enrolado por um pedaço de couro ou pergaminho, era utilizado para escrever ou revelar uma mensagem.

Na criptografia por substituição, um texto é codificado, trocando as letras do mesmo por outras de acordo com um alfabeto cifrado. Ou seja, a posição das letras é mantida, mas seus valores alterados (por exemplo, a letra "A" no texto cifrado pode representar a letra "H" no texto original). Uma das artes do *Kama-sutra*, datado do séc. IV a.C., descreve uma técnica de substituição, na qual cada letra do alfabeto é substituída por seu par.

A segurança de uma mensagem criptografada estava garantida, desde que ninguém (além do remetente e destinatário) soubesse a chave (alfabeto cifrado) utilizada. No entanto, por volta do séc. IX, os árabes inventaram a criptoanálise. O método mais comum de criptoanálise, desenvolvido por al-Kindi, é a análise por freqüência, a qual descobre a chave de uma cifra por substituição, explorando a freqüência de caracteres da mensagem cifrada, comparando com padrões ocorridos em textos na língua original.

A Europa, presa à Idade Média, não evoluiu até o séc. XIII, quando com o Renasci-

mento, a criptografia começou a ser difundida (principalmente entre os cientistas). A partir de então, com o conhecimento necessário e a motivação pela comunicação secreta, novas evoluções das cifras de substituição surgiram, culminando na cifra de substituição polialfabética no séc. XVI. Assim, uma mensagem passou a ser criptografada utilizando dois ou mais alfabetos cifrados. Com isto, cada letra do criptograma podia ter mais de uma representação no alfabeto original. O francês Blaise de Vigenère formulou uma das mais conhecidas cifras da época, a cifra de Vigenère. As cifras polialfabéticas eram consideradas indecifráveis, até que três séculos depois o inglês Charles Babbage e, um pouco depois Friedrick Kasiski, conseguiram quebrá-las.

O período da 1ª Guerra até o final da 2ª Guerra Mundial foi de grande evolução tanto para os criptógrafos quanto para os criptoanalistas, pois os militares precisavam mais do que nunca de uma comunicação segura e descobrir os planos dos seus inimigos. Com a procura constante por cifras mais seguras, o período foi marcado pela mecanização da criptografia e da criptoanálise. Ainda em 1918, o alemão Scherbius inventou a *Enigma*, máquina utilizada cifrar e decifrar uma mensagem, utilizando uma cifra polialfabética. O processo de cifragem é diferente do criado por Vigenère, não ocorrendo a repetição dos padrões detectados pelo método de Babbage. A quebra da *Enigma*, por Rejewski, ocorreu antes da 2ª Guerra e seu processo também foi mecanizado por um conjunto de máquinas chamadas *Bombas*.

Com o início da 2ª Guerra, a *Enigma* começou a sofrer alterações, para aumentar sua segurança, forçando também as evoluções das *Bombas*. E assim, sempre que outros métodos ou máquinas de cifragem eram desenvolvidos, os criptoanalistas detectavam suas falhas e criavam máquinas capazes de decifrar as mensagens. Até que, os alemães inventaram a cifra *Lorenz*, que por sua complexidade não podia ser quebrada pelas *Bombas*. E em 1943, Max Newman desenvolveu o *Colossus*, uma máquina flexível que se adaptava ao problema proposto, sendo a precursora dos computadores modernos.

Com o advento dos computadores e a difusão da criptografia também na sociedade civil, tornou-se necessária a criação de um novo padrão para cifrar mensagens. Neste contexto, os sistemas criptográficos deveriam ser de conhecimento público, possibilitando que a segurança dos mesmos seja questionada e certificada. O conhecimento público de uma cifra está de acordo com o princípio de Auguste Kerckhoffs ("*La Criptographie Militaire*", 1873): "a segurança de um sistema criptográfico deve residir unicamente no segredo da chave, e não no sigilo do algoritmo" (KAHN, 1967).

Em 1974, a IBM apresentou ao *National Bureau of Standards* (NBS) o DES (*Data Encryption Standard*), que se tornou o padrão de criptografia de dados. Apenas em 1998, o NIST (antigo NBS) promoveu uma competição para a escolha de um novo padrão. E em 2001 o algoritmo Rijndael foi eleito como o *Advanced Encryption Standard* (AES).

Todas as técnicas desenvolvidas geravam o mesmo problema, necessitavam que as chaves fossem distribuídas com segurança. Essa distribuição se tornou impraticável à medida que aumentava o número de mensagens cifradas. Uma grande revolução da criptografia na atualidade foi então o desenvolvimento, por Rivest, Shamir e Adleman, da cifra RSA, uma cifra assimétrica que é baseada no conceito da chave pública, de Diffie, Hellman e Merkle. Na cifra RSA, uma chave pública, que pode ser livremente distribuída, depende do produto de dois números primos quaisquer e é utilizada, por uma função, na cifragem de uma mensagem. Essa função é reversível caso se conheçam esses números primos, os quais compõem a chave privada.

2.2 CRIPTOGRAFIA CLÁSSICA

Atualmente, a criptografia clássica, caracterizada pelas cifras desenvolvidas antes do advento dos computadores, não é considerada de uso prático. Isso se deve aos avanços em computação e em criptoanálise que as tornam inseguras (GEBBIE, 2002). Entretanto, a observação de como essas cifras são projetadas e criptoanalisadas pode ser útil no estudo dos sistemas criptográficos modernos (RUBIN, 1997).

Cifras de substituição são caracterizadas pela troca dos símbolos de um texto em claro por outros símbolos, por algum algoritmo de forma predefinida. Um sistema criptográfico **monoalfabético** estabelece um mapeamento único entre cada símbolo de um alfabeto original por um símbolo diferente do alfabeto cifrado (MORAES, 2004). A chave de uma cifra monoalfabética é o alfabeto cifrado, que é, normalmente, uma permutação do alfabeto original.

(GEBBIE, 2002) define a cifra de substituição monoalfabética como:

- $\rho = \gamma = \mathbb{Z}_{26}$;
- κ consiste em todas as permutações possíveis dos 26 símbolos do alfabeto original;
- para cada $k \in \kappa$, com k^{-1} o inverso de k, tem-se: $e_k(x) = k(x)$ e $d_k(y) = k^{-1}(y)$.

Neste caso, considerou-se como alfabeto original as 26 letras do alfabeto latino. Sendo necessário apenas um mapeamento, conforme apresentado na TAB. 2.1. É importante lembrar que em Álgebra, dado n positivo e inteiro, o conjunto $\mathbb{Z}_n = \{0, 1, ..., n-1\}$ e a operação adição(subtração) em $\mathbb{Z}_n$ é definida como a adição(subtração) usual módulo n.

TAB. 2.1: Correspondência numérica com as letras do alfabeto

A	00	H	07	O	14	U	20
B	01	I	08	P	15	V	21
C	02	J	09	Q	16	W	22
D	03	K	10	R	17	X	23
E	04	L	11	S	18	Y	24
F	05	M	12	T	19	Z	25
G	06	N	13				

A principal fraqueza deste tipo de cifra é a preservação da distribuição de freqüências nos criptogramas gerados, tornando-o vulnerável a ataques baseados em freqüências. As freqüências dos caracteres e das seqüências de caracteres em um criptograma possuem uma distribuição semelhante à da língua original (DENNING, 1982).

O ataque a cifras monoalfabéticas é relativamente simples. Por exemplo, se em um criptograma possuir uma alta freqüência do trigrama "XHI", provavelmente a seqüência equivalente no texto em claro deve ser "QUE" (que também possui freqüência elevada), caso a língua original seja o português (PORTUGUÊS, 2002). Então, a letra "X" é equivalente ao "Q" no alfabeto original, assim como o "H" ao "U" e o "I" ao "E". O processo continua até que se consiga decifrar toda a mensagem.

2.2.1 CRIPTOGRAFIA POLIALFABÉTICA

Cifras que produzem criptogramas com uma distribuição mais uniforme de símbolos são, em geral, mais seguras. Uma cifra de substituição polialfabética é uma combinação de cifras monoalfabéticas. Assim, existem várias "chaves monoalfabéticas" e a escolha pela chave a ser usada para cifrar uma letra varia em função da sua posição no texto em claro (MORAES, 2004).

O número de cifras monoalfabéticas é conhecido como **tamanho** ou **período** da chave (m). Na cifragem, uma mensagem é dividida em blocos de tamanho m e cada um dos m caracteres de um bloco é cifrado, alternadamente, usando uma chave monoalfabética diferente. Esse processo suaviza a distribuição dos símbolos, pois cada caractere do texto

em claro pode ter diferentes representações no criptograma.

Formalmente:

- $\rho = \gamma = (\mathbb{Z}_{26})^m$;
- uma chave $k = (k_1, k_2, ..., k_m)$ é composta por m chaves monoalfabéticas devidamente ordenadas;
- para cada $k \in \kappa$ tem-se:
  $e_k(x_1, x_2, ..., x_m) = (e_{k_1}(x_1), e_{k_2}(x_2), ..., e_{k_m}(x_m))$ e
  $d_k(y_1, y_2, ..., y_m) = (d_{k_1}(y_1), d_{k_2}(y_2), ..., d_{k_m}(y_m))$.

2.2.1.1 CIFRA DE VIGENÈRE

A cifra de Vigenère é um caso particular de cifra polialfabética. A simplicidade e usabilidade desta cifra consistem no armazenamento da chave, que é uma simples palavra $k = (k_1, k_2, ..., k_m)$ de tamanho m. Cada letra k_i de uma **palavra chave** define o deslocamento a ser realizado durante a cifragem(decifragem). Assim, $e_{k_i}(x_i) = x_i + k_i$ e $d_{k_i}(y_i) = y_i - k_i$.

De modo geral, a cifra de Vigenère é de definida (LINK, 2003):

- $\rho = \gamma = \kappa = (\mathbb{Z}_{26})^m$;
- para cada $k = (k_1, k_2, ..., k_m) \in \kappa$ tem-se:
  $e_k(x_1, x_2, ..., x_m) = (x_1 + k_1, x_2 + k_2, ..., x_m + k_m)$ e
  $d_k(y_1, y_2, ..., y_m) = (y_1 - k_1, y_2 - k_2, ..., y_m - k_m)$.

Lembre-se que as operações são realizadas em $\mathbb{Z}_{26}$. A cifra de Vigenère foi escolhida para ser utilizada nos experimentos envolvendo sistemas polialfabéticos no decorrer desta dissertação.

2.2.2 CRIPTOANÁLISE POLIALFABÉTICA

Em princípio, por gerar uma distribuição mais uniforme de símbolos, as cifras polialfabéticas são imunes a ataques baseados em analise de freqüências. Todavia, é evidente que um criptograma obtido por uma cifra polialfabética tem propriedades similares ao obtido por uma cifra monoalfabética.

Neste contexto, caso o tamanho da chave seja conhecido, uma cifra polialfabética torna-se vulnerável aos mesmos ataques das cifras monoalfabéticas. Por exemplo, caso o tamanho da chave seja três, os símbolos ocupando a primeira, quarta, sétima, etc. posições de um criptograma foram cifrados utilizando a mesma cifra monoalfabética (ou alfabeto cifrado). Assim, sabendo que o período da chave é três, o problema é reduzido a criptoanalisar três cifras de substituição monoalfabéticas (DENNING, 1982).

Pode-se, então, descrever a criptoanálise de um cifra substituição polialfabética em três etapas:

- determinar o período da chave;
- separar o criptograma em várias partes, uma para cada alfabeto cifrado;
- resolver cada parte como uma cifra monoalfabética, usando análise de freqüências.

Existem alguns métodos para estimar o tamanho da chave usado na cifragem de uma mensagem. Esses métodos serão usados como comparação para o procedimento, baseado em técnicas de recuperação de informações, proposto durante este trabalho. Os métodos mais conhecidos na comunidade são descritos a seguir: **Kasiski** e **Índice de Coincidência**.

2.2.2.1 KASISKI

O método Kasiski foi descrito pela primeira vez em 1863 por Friedrick Kasiski em "*Die Geheimschriften und die Dechiffrierkunst*" (escrita secreta e a arte da decifragem). O método Kasiski é baseado na busca por padrões (seqüências de símbolos) repetidos em um criptograma. Por exemplo, é bem possível que o trigrama "QUE" (muito comum na língua portuguesa) seja substituído freqüentemente pelo mesmo trigrama em uma cifragem polialfabética. Isto porque, em um texto suficientemente grande, os alfabetos cifrados são repetidos constantemente, permitindo que em algumas ocasiões uma mesma seqüência de caracteres seja cifrada pelos mesmos alfabetos.

A idéia geral do método é encontrar duas seqüências de letras idênticas em um criptograma e a distância entre elas na mensagem é um múltiplo do provável período da chave. Assim, o método Kasiski é executado nas seguintes etapas (DENNING, 1982):

- identificar padrões repetidos de três ou mais letras;

- para cada seqüência, anotar as posições iniciais de cada ocorrência dela no criptograma;
- calcular as diferenças entre as posições iniciais de seqüências idênticas;
- determinar todos os fatores das diferenças;
- o tamanho da chave é determinado pelo máximo divisor comum das diferenças encontradas.

Para ilustrar o funcionamento deste método, um trecho da música "Garota de Ipanema" do Tom Jobim foi cifrado utilizando a cifra de Vigenère e "CARLOS" como chave (FIG. 2.2).

```
Chave:       CARL OSC ARLOS CARL OSCAR LOSC ARLOS CA RLOSC A RLO S
Mensagem:    olha que coisa mais linda mais cheia de graca e ela a
Criptograma: QLYL EMG CFTGS OAZD ZAPDR XOAU CYPWS FE XCOUC E VWO S

Chave:       CARLOS CAR LOS C ARL OSCAR LOS CARL OSCARLO S CARLOSC
Mensagem:    menina que vem e que passa seu doce balanco a caminho
Criptograma: OEETBS SUV GSE G QLP DSUSR DSM FOTP PSNAENC S EADTBZQ

Chave:       AR LOS CARL OS CARLO SCARLOS CA RLO SC ARLOSCA R LOS
Mensagem:    do mar moca do corpo dourado do sol de ipanema o seu
Criptograma: DF XOJ OOTL RG EOIAC VQUILRG FO JZZ VG IGLBWOA F DSM

Chave:       CARLOSCAR L OSCA RLO SC ARLOS C A RLOSC ARLO SCARL OSC
Mensagem:    balancado e mais que um poema e a coisa mais linda que
Criptograma: DACLBUCDF P ASKS HFS MO PFPAS G A TZWKC MRTG DKNUL EMG

Chave:       AR LO SC ARLOSC
Mensagem:    eu já vi passar
Criptograma: EL UO NK PRDGST
```

FIG. 2.2: Exemplo para mostrar o funcionamento do método Kasiski

Para o exemplo acima, foram analisadas todas as seqüências três letras. Todos os trigramas repetidos foram sublinhados e suas posições extraídas. Observe que os espaços e acentos foram ignorados. A TAB. 2.2 apresenta as distâncias entre as seqüências e seus fatores.

TAB. 2.2: Distâncias entre trigramas repetidos

Seqüência	**Posições**	**Diferenças e fatores**
LEM	4, 166	$162 = 2 * 3^4$
EMG	5, 167	$162 = 2 * 3^4$
DSM	64, 124	$60 = 2^2 * 3 * 5$
LRG	94, 106	$12 = 2^2 * 3$
FPA	135, 147	$12 = 2^2 * 3$
PAS	136, 148	$12 = 2^2 * 3$

Como determinado pela fatoração em números primos, o maior fator comum (ou máximo divisor comum) das distâncias entre esses trigramas é 2 * 3 = 6. Assim, o tamanho da chave deste criptograma foi encontrado corretamente pelo método Kasiski.

2.2.2.2 ÍNDICE DE COINCIDÊNCIA

Em 1920, William Friedman publicou "*The Index of Coincidence and Its Application in Cryptography*", em que descreve o Índice de Coincidência (IC) e como ele pode ser aplicado em criptoanálise polialfabética. O IC nada mais é do que a probabilidade de dois símbolos de uma mensagem, tomados ao acaso, corresponderem à mesma letra (DENNING, 1982). Assim,

$$IC(m) = \frac{\sum_A^Z n_i(n_i - 1)}{n(n-1)},$$

em que n_i é a freqüência de ocorrência da letra i na mensagem m e n é o tamanho desta mensagem.

Exemplificando, pode-se calcular o IC do trecho da música de Tom Jobim criptografado anteriormente. Assim, sabendo que o criptograma é composto por 181 letras e com base na distribuição de freqüências do mesmo (TAB. 2.3), obteve-se o IC aproximadamente igual a 0,046.

De forma genérica, o IC pode ser calculado como:

$$IC(m) = \sum_A^Z p_i^2,$$

em que p_i é a probabilidade de ocorrência da letra i na mensagem m. Neste contexto, observando a distribuição de freqüências de uma língua qualquer é possível determinar o Índice de Coincidência desta língua (IC_L). Para o português, de acordo com uma tabela de freqüências elaborada por (PORTUGUÊS, 2002) (TAB. 2.4), o Índice de Coincidência (IC_P) obtido foi aproximadamente 0,078.

TAB. 2.3: Freqüências das letras do criptograma

Letra	**Freqüência**	**Letra**	**Freqüência**	**Letra**	**Freqüência**
A	11	J	02	S	17
B	04	K	04	T	08
C	09	L	09	U	08
D	11	M	06	V	04
E	11	N	04	W	04
F	09	O	14	X	03
G	12	P	09	Y	02
H	01	Q	04	Z	06
I	03	R	06		

TAB. 2.4: Freqüências da ocorrência de letras no português

Letra	**Freqüência (%)**	**Letra**	**Freqüência (%)**	**Letra**	**Freqüência (%)**
A	14,63	J	0,4	S	7,81
B	1,04	K	0,02	T	4,34
C	3,88	L	2,78	U	4,63
D	4,99	M	4,74	V	1,67
E	12,57	N	5,05	W	0,01
F	1,02	O	10,73	X	0,21
G	1,3	P	2,52	Y	0,01
H	1,28	Q	1,2	Z	0,47
I	6,18	R	6,53		

Similarmente, pode-se calcular o Índice de Coincidência para um texto completamente aleatório (IC_R). Para um alfabeto composto por 26 letras (caso do português), o índice calculado é aproximadamente 0,038.

Então, fica claro que o Índice de Coincidência está intimamente relacionado com o alfabeto e a linguagem em questão. É evidente também que o IC de um texto em claro é igual ao criptograma do mesmo gerado por uma cifra monoalfabética. Neste caso, as probabilidades individuais são trocadas, mas seu somatório permanece inalterado.

Friedman propôs, ainda, uma forma de estimar o período de uma chave de um criptograma. O Índice de Coincidência de um criptograma, cifrado por uma cifra polialfabética, deve variar entre o IC_L e o IC_R da linguagem em questão, dependendo do tamanho da chave usada. Conseqüentemente, o tamanho da chave t pode ser determinado, em função do IC de um criptograma m, por:

$$t \approx \frac{n * (IC_L - IC_R)}{(n-1) * IC(m) - nIC_R + IC_L}$$

Para o exemplo aqui descrito, o período da chave encontrado, por meio do IC, foi cinco.

Embora o tamanho de chave real seja seis, o resultado pode ser considerado satisfatório, mesmo porque o erro no período da chave foi apenas de um e o texto utilizado era de um tamanho relativamente pequeno.

Uma grande desvantagem do IC é que as diferenças entre os ICs de chaves de tamanhos vizinhos diminuem com o aumento do período da chave, dificultando seu uso para determinar o tamanho da chave de um criptograma. Entretanto, o IC pode ser usado juntamente com o método Kasiski para fortalecer o processo de criptoanálise.

Existe, ainda, em criptoanálise polialfabética a teoria de que quanto mais próximo for o período da chave do tamanho do texto, mais difícil é a criptoanálise; e que se os dois forem do mesmo tamanho, ela se torna impraticável (SINGH, 2001).

2.3 CIFRAS DE BLOCOS

Em virtude do atual desuso das cifras clássicas, é necessário também estudar algoritmos contemporâneos de criptografia. As cifras de blocos são bem mais seguras em face do poder computacional que atualmente pode ser utilizado para criptoanálise.

Um cifra de bloco (MENEZES, 1996) é uma função $e : \mathbb{Z}_2^m \times \kappa \rightarrow \mathbb{Z}_2^m$, que converte um texto em claro p de m bits em um criptograma c de mesmo tamanho, de acordo com um mapeamento $e_k(p)$ definido por uma chave $k \in \kappa$. Para permitir uma decifragem única, a função de cifragem de uma chave fixa deve ser bijetiva. A função de decifragem $d_k(c)$ é o mapeamento inverso.

Como pôde ser observado anteriormente, a cifragem é realizada com blocos de um tamanho fixo pré-definido. As mensagens, em prática, têm normalmente tamanhos consideravelmente maiores do que o tamanho do bloco. Os **modos de operação** são utilizados para permitir a cifragem de textos com os mais variados tamanhos, descrevendo a divisão da mensagem em blocos e a maneira como os mesmos são cifrados. Existem alguns desses modos, sendo o ECB (*electronic codebook*) o foco deste estudo.

No modo ECB, um texto em claro é dividido em um número inteiro de blocos e cada um é cifrado (ou decifrado) independentemente de acordo com o algoritmo e a chave escolhida. O criptograma é obtido pela concatenação dos blocos cifrados. Assim, $c_i = e_k(p_i)$ e $p_i = d_k(c_i)$, em que p_i e c_i representam o bloco i do texto em claro e do criptograma, respectivamente.

Normalmente, um algoritmo de enchimento (*padding*) é utilizado para garantir que o tamanho de uma mensagem seja múltiplo do tamanho do bloco. Em (FIPS 81, 1980),

outros modos de operação são descritos, como o CBC (*Cipher Block Chaining*), em que a saída da cifragem de um bloco é usada na cifragem do bloco seguinte.

Uma desvantagem do modo ECB é que blocos de textos em claro idênticos (cifrados com uma mesma chave) geram blocos cifrados iguais. Este método, apesar de não acrescentar nada à segurança de uma cifra, continua sendo utilizado, principalmente no estudo da criptoanálise.

A repetição de blocos cifrados não ocorre na cifragem em modo CBC. Entretanto, nesse método, um eventual erro na transferência de um dos blocos comprometeria a decifragem de dois blocos. Além disso, usando o método ECB, é possível facilmente, em virtude da independência entre os blocos, paralelizar o processo de cifragem.

Assim, podem-se considerar as cifras operadas em modo ECB um caso particular de um sistema de substituição monoalfabética, considerando como alfabeto os 2^m símbolos possíveis para um bloco de tamanho m.

As tabelas de domínio e imagem são tão grandes que inviabilizam o uso próspero dos ataques criptoanalíticos tradicionais baseados em freqüências. Entretanto, as repetições de blocos podem ser identificadas por meio das técnicas de RI, caracterizando uma fraqueza dessas cifras.

O DES e o AES são as cifras de blocos mais conhecidas, especialmente porque são oficialmente certificadas como padrões de criptografia. Eles possuem tal *status* com base no fato de que o texto em claro e a chave são suficientemente misturados de forma que os testes de validação existentes não conseguem identificar qualquer correlação entre o criptograma gerado e a chave ou mensagem original.

Como será observado posteriormente, o DES e o AES são algoritmos que consistem em várias iterações, que misturam os bits da mensagem com os da chave, usando operações de permutação e de substituição. O objetivo é misturar os dados completamente de forma que cada bit do criptograma dependa de todos os bits da mensagem e da chave (LAMBERT, 2004).

2.3.1 DES

O DES foi um padrão mundial durante 20 anos e, apesar de mostrar os sinais de sua idade, tem resistido às mais diversas formas de criptoanálise. O DES é um algoritmo simétrico de criptografia que cifra blocos de **64 bits**. O DES é baseado na estrutura de Feistel (FEISTEL, 1973).

A idéia original de Feistel é dividir um bloco de comprimento m (par) em dois sub-blocos de comprimento $m/2$. Sejam $B_i = R_i|L_i$ e k_i, respectivamente, o bloco e a sub-chave da *i-ésima* interação de um cifrador (de um total de r iterações) que utiliza uma função f. A cifragem de uma interação (FIG. 2.3) é determinada, pela saída da iteração anterior, por:

$$R_i = L_{i-1} \oplus f(R_{i-1}, k_i)$$

$$L_i = R_{i-1}$$

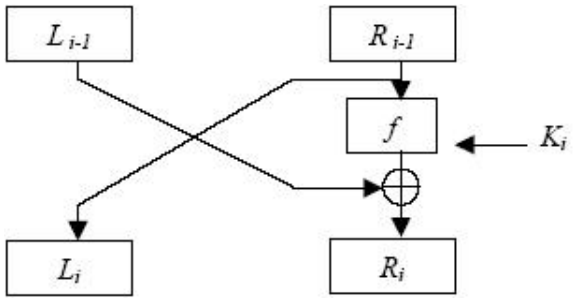

FIG. 2.3: Estrutura de Feistel para cifrar $(i = 1, 2, ..., r)$

Isso é uma grande idéia, pois garante a reversibilidade do processo. A decifragem é possível devido à seguinte propriedade da soma módulo dois (operação **ou-exclusivo**): $(X \oplus Y) \oplus Y = X$, em que X e Y são blocos de mesmo tamanho. Assim, cada iteração da decifração (FIG. 2.4) é definida por:

$$R_{i-1} = L_i$$

$$L_{i-1} = R_i \oplus f(R_{i-1}, k_i)$$

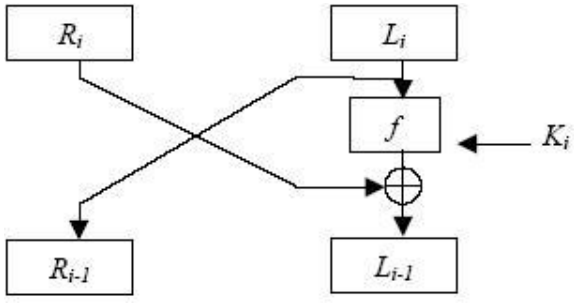

FIG. 2.4: Estrutura de Feistel para decifrar $(i = r, r-1, ..., 1)$

A cifragem do DES é realizada em 16 rodadas (ou iterações). A chave utilizada é de **64 bits**, que efetivamente tornam-se 56 bits após a exclusão dos bits de paridade. De forma básica, o algoritmo nada mais é do que a aplicação de duas técnicas milenares de cifração: substituição e permutação, conhecidas hoje como princípios de **confusão** e **difusão**, respectivamente.

O esquema da FIG. 2.5 indica que os 64 bits de um bloco são submetidos a uma **permutação inicial**. O bloco então é divido ao meio e as 16 interações são realizadas. Finalmente, as duas metades são concatenadas e o bloco cifrado é gerado após uma **permutação final**.

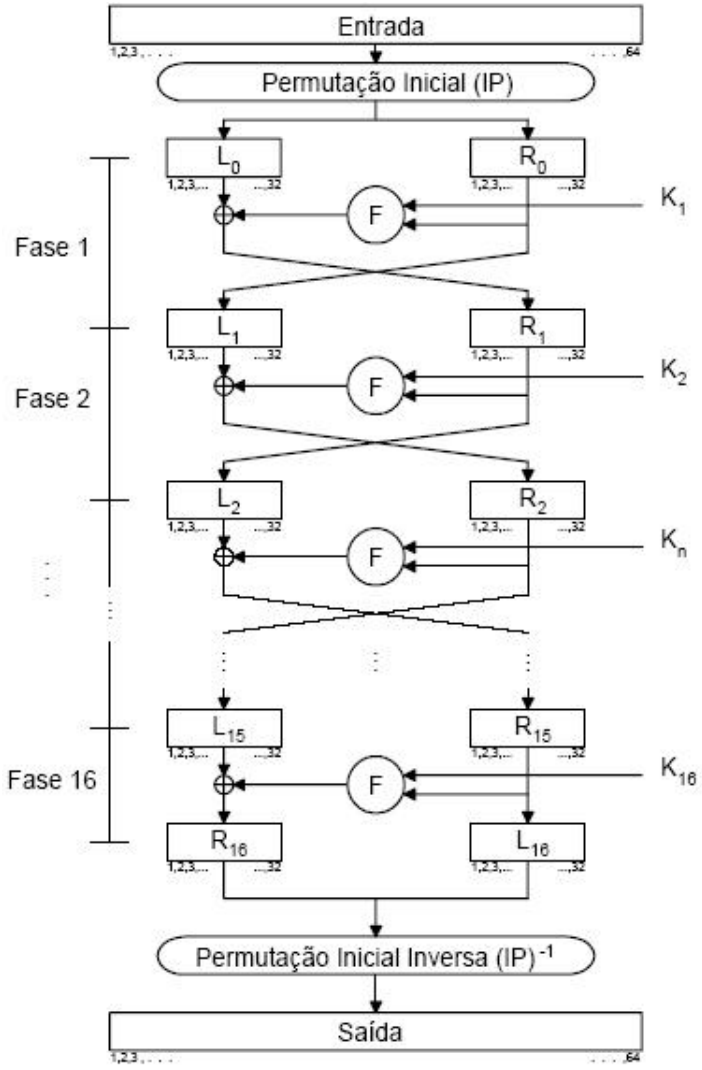

FIG. 2.5: Fluxograma do DES

A segurança do DES reside na função f utilizada. Ela pode ser definida por $f(R_{i-1}, k_i) =$

$P(S(E(R_{i-1}) \oplus k_i))$, em que:

- $E(X)$ é uma função de expansão que converte uma entrada de 32 bits em 48 bits;
- $S(X)$ é uma função de substituição, realizada por meio de 8 tabelas (**caixas-S**), que substitui os 48 bits de entrada por uma saída de 32 bits;
- $P(X)$ é uma função para permutação de um bloco de 32 bits.

Em (FIPS 46-3, 1999) são definidas as permutações inicial e final; as funções $E(X)$, $S(X)$ e $P(X)$; além do gerador das 16 sub-chaves k_i de 48 bits utilizadas pelo DES.

2.3.2 AES

O algoritmo de Rijndael, eleito em 2001 como novo padrão de criptografia (AES), é uma cifra simétrica de bloco, que processa dados de 128 bits, usando uma chave de 128, 192 ou 256 bits. O número de iterações é 10, 12 ou 14 para as chaves de 128, 192 ou 256 bits respectivamente.

O esquema da FIG. 2.6 mostra que cada rodada é composta por quatro diferentes transformações: *SubByte*, *ShiftRow*, *MixColumn* e *AddRoundKey*. A última rodada é diferenciada pela não utilização da função *MixColumn*. Além disso, existe um passo "zero" com a transformação *AddRoundKey*. As sub-chaves são extraídas de uma função conhecida como *KeyExpansion*, descrita em (FIPS 197, 2001).

SubByte é uma transformação não-linear, que opera sobre cada byte de forma independente, usando uma tabela de substituição (TAB. 2.5). A seleção se processa da seguinte forma: sendo, a entrada, um byte em hexadecimal representado por dois dígitos na base 16 ($x_{16}|y_{16}$), obtém como resultado o byte localizado na intersecção da linha x_{16} com a coluna y_{16}.

Na transformação de permutação *ShiftRow*, os bytes de entrada ($b_0 b_1 ... b_{15}$) trocam suas posições. Na saída os bytes ficam assim ordenados: $b_0 b_5 b_{10} b_{15} b_4 b_9 b_{14} b_3 b_8 b_{13} b_2 b_7 b_{12} b_1 b_6 b_{11}$.

A transformação linear *MixColumn* consiste em uma multiplicação de matrizes. Os bits de entrada são divididos em quatro palavras de 32 bits. Cada palavra é então multiplicada pela matriz M_{32x32} de valores binários (FIG. 2.7). Os 128 bits de saída são obtidos pela concatenação das quatro palavras resultantes das multiplicações realizadas.

Por fim, a operação *AddRoundKey* é uma simples soma módulo dois da saída da transformação anterior com a sub-chave da respectiva rodada.

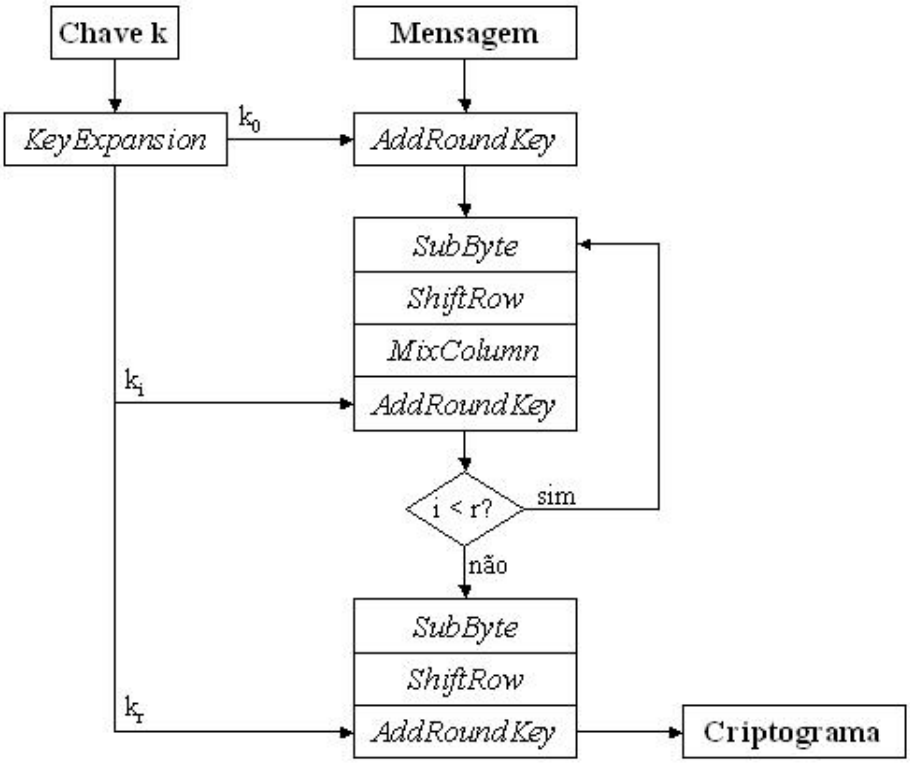

FIG. 2.6: Fluxograma do AES

Outra característica importante é que o processo para cifrar e decifrar não é o mesmo como na maioria das cifras de bloco. O AES não utiliza uma estrutura de Feistel. Assim, na decifragem, a obtenção dos valores originais se deve ao uso das operações inversas das aqui descritas. As transformações inversas e outras informações sobre o AES são encontradas em (FIPS 197, 2001).

2.3.3 ATAQUES A CIFRAS DE BLOCOS

O objetivo da criptoanálise é obtenção da chave de cifragem, ou ainda do texto original sem o conhecimento da chave. Caracterizam-se também como criptoanálise os processos que, de alguma forma, descobrem fraquezas das cifras e contribuem para atingir o objetivo final.

Neste contexto, várias técnicas de criptoanálise de cifras de blocos vêm sendo desenvolvidas. Alguns trabalhos recentes (BIRYUKOV, 2004) (STANDAERT, 2003) (WAGNER, 2004) mostram uma evolução dessas técnicas. Embora nenhum ataque baseado em procedimentos lingüísticos tenha sido encontrado, é relevante investigar e avaliar a

TAB. 2.5: Tabela da operação *SubByte* do AES

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	63	7C	77	7B	F2	6B	6F	C5	30	01	67	2B	FE	D7	AB	76
1	CA	82	C9	7D	FA	59	47	F0	AD	D4	A2	AF	9C	A4	72	C0
2	B7	FD	93	26	36	3F	F7	CC	34	A5	E5	F1	71	D8	31	15
3	04	C7	23	C3	18	96	05	9A	07	12	80	E2	EB	27	B2	75
4	09	83	2C	1A	1B	6E	5A	A0	52	3B	D6	B3	29	E3	2F	84
5	53	D1	00	ED	20	FC	B1	5B	6ª	CB	BE	39	4ª	4C	58	CF
6	D0	EF	AA	FB	43	4D	33	85	45	F9	02	7F	50	3C	9F	A8
7	51	A3	40	8F	92	9D	38	F5	BC	B6	DA	21	10	FF	F3	D2
8	CD	0C	13	EC	5F	97	44	17	C4	A7	7E	3D	64	5D	19	73
9	60	81	13	DC	22	2ª	90	88	46	EE	B8	14	DE	5E	0B	DB
A	E0	32	4F	0A	49	06	24	5C	C2	D3	AC	62	91	95	E4	79
B	E7	C8	3A	6D	8D	D5	4E	A9	6C	56	F4	EA	65	7A	AE	08
C	BA	78	37	2E	1C	A6	B4	C6	E8	DD	74	1F	4B	BD	8B	8A
D	70	3E	B5	66	48	03	F6	0E	61	35	57	B9	86	C1	1D	9E
E	E1	F8	98	11	69	D9	8E	94	9B	1E	87	E9	CE	55	28	DF
F	8C	A1	89	0D	BF	E6	42	68	41	99	2D	0F	B0	54	BB	16

aplicação de técnicas de RI.

Essas técnicas podem ser utilizadas na descoberta e agrupamento de características de criptogramas, de maneira a revelar padrões que auxiliem no processo da criptoanálise. Procedimentos de identificação de emissor e de detecção de mudanças de chave, também podem ser desenvolvidos utilizando as técnicas de recuperação de informações.

A obtenção de uma chave pode ser feita por meio busca entre todas as chaves possíveis. Este tipo de ataque é chamado de **força bruta** e mede o limite superior natural da segurança de uma cifra (BIHAM, 1993). Assim, todo e qualquer ataque desenvolvido tem como objetivo reduzir o tempo de criptoanálise em relação à força bruta.

Observa-se, então, que essa dissertação apresenta uma abordagem inovadora para o estudo da criptoanálise em cifras de blocos. A localização de padrões em criptogramas questiona a inexistência da correlação entre os dados de entrada e de saída de uma cifra.

Nos ataques, normalmente, são determinados alguns dos bits da chave, e os bits restantes são obtidos pela força bruta. Existe a possibilidade de integrar o uso de técnicas de RI aos ataques tradicionais de maneira a evitar o trabalho exaustivo da força bruta, resultando em uma maior eficiência da criptoanálise.

Os ataques são normalmente classificados, de acordo com as necessidades e objetivos dos criptoanalistas, em (MENEZES, 1996):

- ataque com texto cifrado: o criptoanalista tenta deduzir a chave de decifragem ou o próprio texto em claro a partir somente do criptograma;

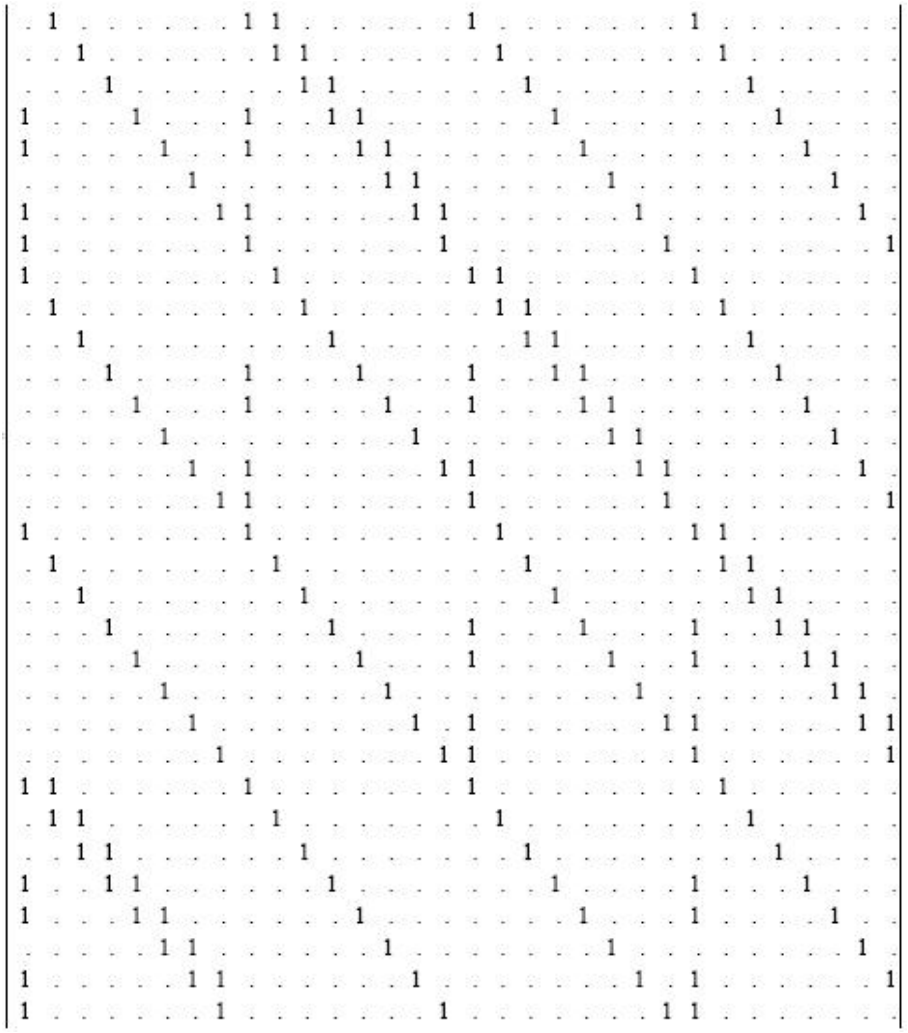

FIG. 2.7: Ilustração da matriz M_{32x32} (Obs: $'.' = 0$)

- ataque com texto em claro conhecido: o criptoanalista tem a quantidade que desejar de textos em claro e seus respectivos criptogramas. Seu objetivo é descobrir a chave ou como decifrar um novo criptograma;

- ataque com texto em claro escolhido: o criptoanalista escolhe os textos a serem cifrados. Os blocos de textos escolhidos são capazes de revelar mais informações sobre a chave;

- ataque com texto em claro escolhido adaptado: os textos em claro escolhidos dependem dos criptogramas recebidos em solicitações anteriores;

- ataque com criptograma escolhido: o criptoanalista escolhe o criptograma, e também

tem acesso ao respectivo texto em claro. Imagine possuir acesso ao equipamento de decifragem (mas não à chave). O objetivo é, então, deduzir o texto em claro de um novo criptograma sem esse equipamento;

- ataque com criptograma escolhido adaptado: os criptogramas escolhidos dependem dos textos em claros obtidos anteriormente.

Nas subseções seguintes são apresentados os métodos de criptoanálise de mais destaque na literatura.

2.3.3.1 CRIPTOANÁLISE DIFERENCIAL

A criptoanálise diferencial, proposta por Eli Biham e Adi Shamir em (BIHAM, 1991), é utilizada em ataques com texto em claro escolhido contra cifras simétricas de bloco. Neste método, a chave pode ser determinada de forma mais eficiente que utilizando a força bruta.

A criptoanálise diferencial é baseada na comparação entre a diferença (ou exclusivo) de dois textos em claro e a diferença dos criptogramas correspondentes. Essas diferenças, produzidas por alterações controladas do texto em claro, podem ser usadas para apontar as probabilidades das chaves possíveis e localizar a chave mais provável.

2.3.3.2 CRIPTOANÁLISE LINEAR

A criptoanálise linear, desenvolvida por Mitsuru Matsui em (MATSUI, 1993), é um tipo de ataque que usa aproximações lineares para descrever a ação de um criptosistema, como o DES. Este ataque é aplicado a textos em claro conhecidos, e quando aplicado ao DES é mais eficiente que a criptoanálise diferencial.

Com base em equações lineares de bits de entrada e saída, podem-se sugerir valores para equações lineares de bits da chave, válidas com alguma probabilidade. Este fato só poderá ser explorado caso a probabilidade p de uma equação linear de bits da chave for diferente de 0,5, e quanto maior $|p - 0,5|$, menor a quantidade de pares (textos em claro e criptograma) necessários para sugerir corretamente a chave.

Um ataque combinando a criptoanálise diferencial e a linear, chamado de criptoanálise diferencial-linear, foi descrito por Suzan Langford e Martin Hellman em (LANGFORD, 1994).

2.3.3.3 ATAQUE DE DAVIES

Em 1987, Davies (DAVIES, 1987) descreveu um ataque para o DES baseado na distribuição não-uniforme das saídas de caixas-S adjacentes. Teoricamente, isto permite a determinação de 16 dos bits da chave.

Este ataque, aplicado a textos em claro conhecidos, foi aprimorado por Eli Biham e Alex Biryukov (BIHAM, 1997), possibilitando a quebra do DES de forma mais rápida que usando a força bruta. Entretanto, este método é menos eficiente do que as criptoanálises diferencial e linear.

2.3.3.4 ATAQUES POR INTERPOLAÇÃO

Thomas Jakobsen e Lars Knudsen (JAKOBSEN, 1997) apresentaram outro ataque a cifras de blocos, o ataque por interpolação. Este método é útil para cifras que usam funções algébricas simples, como as caixas-S. Este ataque, aplicado a textos em claro escolhidos, é baseado na fórmula de interpolação de Lagrange. Além disso, cifras provadamente seguras às criptoanálises diferencial e linear podem ser vulneráveis ao mesmo.

3 RECUPERAÇÃO DE INFORMAÇÕES

A linguagem é um dos aspectos fundamentais do comportamento humano, pois permite a interação entre os indivíduos. Assim, a lingüística computacional procura aproximar o computador da realidade do homem, por meio de ferramentas que proporcionem uma comunicação mais natural.

Uma das principais necessidades do homem é a localização e recuperação das informações que lhe sejam convenientes. Assim, surgiram os sistemas de recuperação de informações. Neles, um usuário pode encontrar os dados desejados sem precisar analisar todas as informações de uma base de dados (WIVES, 1997). Somente informações textuais serão utilizadas neste trabalho, mas imagens, sons e vídeos também são considerados fontes de dados.

Alguns sistemas de recuperação de informações têm dependido fortemente do modelo "saco de palavras" para a representação de documentos, negligenciando qualquer conhecimento lingüístico da linguagem em questão, mas são altamente influenciados por características estatísticas da linguagem. Muitas tarefas de RI são resolvidas pela busca de textos semelhantes, mesmo que não se saiba o conteúdo dos mesmos.

Este contexto, juntamente com a perspectiva de que as características lingüísticas são propagadas nos textos cifrados, proporcionou o uso de técnicas de RI em criptoanálise. A premissa básica é que um criptograma é um documento normal escrito em uma língua desconhecida.

Uma larga variedade de técnicas de RI foi desenvolvida ao longo dos anos. Neste capítulo são descritas apenas aquelas que foram estudadas e aplicadas no âmbito deste trabalho. Primeiramente é necessário o processamento de uma coleção de textos (**corpus**). O modelo de espaço vetorial foi usado para representar o corpus que será utilizado pelos sistemas de agrupamento e classificação.

3.1 PROCESSAMENTO DE TEXTOS

Muitas tecnologias associadas ao processamento de textos utilizam descrições e representações lingüísticas em vários níveis, como o nível sintático, o semântico ou o discursivo (GASPERIN, 2000).

Entretanto, na criptoanálise, trabalha-se com textos cifrados, em que não é levada em consideração a estrutura da linguagem. Assim, torna-se dispensável a análise sintática e semântica dos documentos, realizando-se apenas o tratamento morfológico dos mesmos.

As palavras são de fundamental importância para um documento, pois elas podem ser usadas para identificar contexto ou assunto desse documento. Dessa forma, a palavra costuma ser o elemento utilizado por um sistema de recuperação. Assim, surge a necessidade de colocar as palavras de um corpus em uma estrutura auxiliar, e a partir delas acessar os textos.

O Arquivo invertido é a estrutura comumente utilizada em sistemas de RI para o armazenamento de um corpus. O processo de mapeamento dos textos para essa estrutura é chamado de indexação. Outros tipos de arquivos podem ser encontrados (VAN RIJSBERGEN, 1979). Entretanto, segundo (SALTON, 1983), esta é uma das estruturas mais eficientes para a indexação de textos.

A estrutura de arquivos invertidos (FRAKES, 1992) é caracterizada por possuir um arquivo de índice, em que cada termo corresponde a uma palavra e é seguido pelos apontadores. Estes, por sua vez, informam em quais documentos existe a palavra, e sua freqüência nos mesmos. As posições de uma palavra no texto podem ser informadas para possibilitar a localização de palavras adjacentes.

O processo de indexação é constituído de algumas etapas que variam dependendo do modelo utilizado (WIVES, 2002). Porém, podem-se destacar as seguintes: identificação das palavras, eliminação de palavras consideradas irrelevantes (*stopwords*), lematização (*word stemming*) e seleção de termos (*feature selection*). A única etapa considerada obrigatória é a de identificação de palavras. A realização das outras depende das necessidades da aplicação.

Na identificação de palavras é feita a análise léxica dos documentos. É neste processo que são eliminados os caracteres inválidos e é feita a filtragem das seqüências de controle (ou de formatação de texto). Pode ainda ser feita uma correção ortográfica ou validação dos termos, caso um dicionário seja utilizado.

Em princípio, nem todas as palavras podem ser incluídas na estrutura de índice, pois algumas não são significativas, caracterizadas por pouco contribuirem semanticamente para o conteúdo dos textos mas usadas apenas para fazer o encadeamento de idéias, como as preposições. Além disso, as palavras muito freqüentes não contribuiem para a discriminação de um texto com relação a outros de uma mesma coleção. Essas palavras,

chamadas de *stopwords*, devem estar numa lista, a *stoplist*, para não serem indexadas.

As *stopwords* aparecem em muitos documentos, e a indexação das mesmas pode comprometer a precisão e a eficiência do sistema. A construção de uma *stoplist* pode ser manual ou ainda automática, identificando como *stopwords* as palavras com maior freqüência documental. A freqüência documental de uma palavra é dada pela quantidade de textos em que a mesma ocorre.

A lematização é um processo que unifica, em um único termo para indexação, um conjunto de palavras de mesma origem morfológica. Um aspecto negativo desse processo são os conflitos gerados por palavras distintas mapeadas para uma mesma palavra básica. Esta etapa é altamente influenciada pela linguagem dos documentos e alguns experimentos com textos da língua inglesa podem ser vistos em (KRAAIJ, 1996).

A seleção de termos pode ser feita para adicionar aos arquivos invertidos apenas as palavras consideradas relevantes. (YANG, 1997) apresenta alguns métodos de seleção e uma comparação entre eles.

As etapas opcionais buscam melhorar a eficiência de sistemas de RI geralmente pela redução do vocabulário de um corpus. Entretanto, especialistas ainda discutem a validade desta afirmação, já que em alguns domínios a realização das mesmas pode comprometer os resultados de um sistema (WIVES, 1997).

É importante lembrar que nosso objetivo principal é identificar padrões em criptogramas, e fica evidente que alguns padrões podem ser, simplesmente, removidos com a realização dessas etapas. Assim, a execução das mesmas torna-se impraticável.

Baseado no corpus da FIG. 3.1, a TAB. 3.1 apresenta o resultado do processo de indexação. Durante este processo, os acentos foram ignorados e toda letra maiúscula foi convertido em minúscula. Para diminuir o vocabulário, foi definida a seguinte *stoplist*: 'a', 'no', 'de', 'ao', 'todos', 'os' e 'muito'. Considerando a utilização de um modelo "saco de palavras" para representar os documentos, não é necessário armazenar a posição da palavra em um texto.

Dessa forma, é possível encontrar facilmente os documentos que contém determinada(s) palavra(s). Entretanto, é parte do trabalho de criptoanálise a localização de padrões em textos cifrados. A chave é o elemento de um sistema criptográfico que determina o "vocabulário" de um criptograma. Assim, textos cifrados com uma mesma chave devem ser mais próximos que os cifrados com chaves distintas (vide Seção 3.2.1).

```
Corpus:
Doc1. André foi à praia no final de semana passado.
Doc2. Eduardo vai ao cinema todos os finais de semana.
Doc3. Fabrício gosta muito de pizza.
Doc4. André joga futebol todos os finais de semana
```

FIG. 3.1: Exemplo de um corpus

TAB. 3.1: Resultado da indexação

Termos	**Apontadores**					
	Doc.	*Freq.*	*Doc.*	*Freq.*	*Doc.*	*Freq.*
andre	1	1	4	1		
foi	1	1				
praia	1	1				
final	1	1				
semana	1	1	2	1	4	1
passado	1	1				
eduardo	2	1				
vai	2	1				
cinema	2	1				
finais	2	1	4	1		
fabricio	3	1				
gosta	3	1				
pizza	3	1				
joga	4	1				
futebol	4	1				

3.2 MODELO DE ESPAÇO VETORIAL

A comparação entre documentos pode ser realizada por um cálculo de similaridade feito a partir dos termos existentes nos mesmos. O modelo de espaço vetorial (MANNING, 1999) é um dos modelos, para representação de documentos, mais utilizados devido à sua simplicidade conceitual e à facilidade para calcular a proximidade entre os textos.

Nesse modelo, desenvolvido por Gerard Salton (SALTON, 1983), os documentos são representados como vetores de termos. Cada um dos termos corresponde a uma palavra específica e possui um **peso** associado, ou o grau de importância desta palavra no texto. É essencial que os vetores gerados estejam normalizados, ou seja, todos os vetores possuam os mesmos termos nas mesmas posições, mudando apenas os valores de seus pesos de acordo com o documento em questão.

A partir de um arquivo de índice, obtido no processo de indexação, o espaço vetorial de um corpus pode ser gerado. A TAB. 3.2 apresenta o espaço vetorial do corpus usado

anteriormente como exemplo. Note que apenas a freqüência do termo em um documento é utilizada como critério de pesagem. O tamanho do vocabulário de um corpus determina a dimensão do espaço.

TAB. 3.2: Espaço vetorial do exemplo utilizado

Termos	Documentos			
	1	*2*	*3*	*4*
andre	1	0	0	1
foi	1	0	0	0
praia	1	0	0	0
final	1	0	0	0
semana	1	1	0	1
passado	1	0	0	0
eduardo	0	1	0	0
vai	0	1	0	0
cinema	0	1	0	0
finais	0	1	0	1
fabricio	0	0	1	0
gosta	0	0	1	0
pizza	0	0	1	0
joga	0	0	0	1
futebol	0	0	0	1

Em princípio a freqüência de uma palavra pode ser uma boa medida de pesagem. Entretanto, esta pode não representar fielmente a importância de um termo. As técnicas mais comuns para o cálculo de relevância (pesagem de um termo) são baseadas nas freqüências dos termos no corpus (WIVES, 2002).

Teoricamente, uma palavra presente em muitos documentos de um corpus deve ser menos importante do que uma de menor freqüência documental. O tamanho do documento também pode ser levado em consideração para que textos pequenos não sejam discriminados. Uma variedade de fórmulas foi desenvolvida para estimar o peso de uma palavra (SALTON, 1987) (MANNING, 1999).

As técnicas de pesagem para o modelo de espaço vetorial são amplamente baseadas nas estatísticas de um simples termo. O uso dessas técnicas está diretamente relacionado à aplicação em questão. Em geral, as cifras procuram gerar criptogramas com uma distribuição mais uniforme das freqüências dos termos. Esses termos são os padrões que serão procurados pelo sistema de RI e dependem do sistema criptográfico em questão. As estatísticas são, então, distorcidas na cifragem, resultando em uma grande dificuldade de avaliar o impacto da pesagem. Assim, para enfatizar a importância da freqüência, o peso foi definido apenas pela freqüência do termo em um documento.

A distância entre documentos indica seu grau de similaridade, ou seja, textos que possuem mais palavras em comum são normalmente inseridos em uma mesma região do espaço (FIG. 3.2) e, em teoria, tratam de assuntos semelhantes.

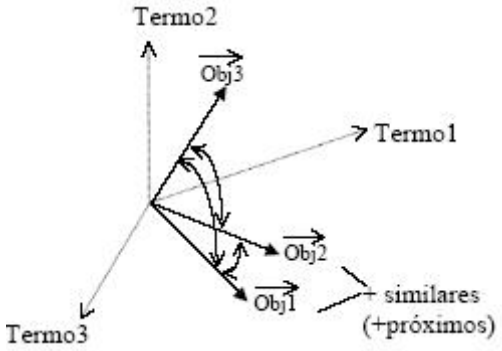

FIG. 3.2: Modelo de espaço vetorial

3.2.1 MATRIZ DE SIMILARIDADES

Com base no espaço vetorial de um corpus, é possível calcular o quão próximos estão dois textos. Existem várias métricas de associação de documentos, como a distância de Euclides, o coeficiente de Dice e o coeficiente de Jaccard (FRAKES, 1992). Entretanto a medida de similaridade que vem se mostrando mais adequada aos sistemas de recuperação de informações é a do cosseno (SALTON, 1987). Aquele mesmo de álgebra linear, obtido pelo produto escalar dos vetores e dividido pelo produto entre os módulos destes.

$$\cos(u, v) = \frac{\sum(u_i * v_i)}{\sqrt{\sum u_i^2 * \sum v_i^2}}$$

Dessa forma, pode-se calcular a similaridade entre todos os documentos, dois a dois, de um corpus, montando-se uma matriz de similaridades (TAB. 3.3). É fácil observar que esta matriz é simétrica ($S_{ij} = S_{ji}$) e que todos os elementos da diagonal principal são iguais a um ($S_{ii} = 1$).

TAB. 3.3: Matriz de similaridades do exemplo utilizado

Documentos	*1*	*2*	*3*	*4*
1	1	0,183	0	0,365
2	0,183	1	0	0,4
3	0	0	1	0
4	0,365	0,4	0	1

3.3 AGRUPAMENTO

Os sistemas de RI podem ser utilizados para solucionar alguns problemas para descoberta de conhecimento. Os problemas mais comuns, encontradas na literatura, são: extração de informações, sumarização, agrupamento (*clustering*) e classificação (WIVES, 2002)

Métodos de agrupamento de documentos têm sido usados na tentativa de agrupar textos de conteúdo similar, mesmo que desconhecido. Em criptografia, uma chave determina a linguagem particular de um sistema criptográfico e assim textos cifrados com uma mesma chave são mais similares. Então, é interessante o uso desses métodos para separar criptogramas de acordo com a chave utilizada.

As técnicas de agrupamento (JAIN, 1988) agrupam itens em grupos baseadas no cálculo do grau de associação entre eles. Os grupos (ou clusters) formados devem ter um alto grau de associação entre seus membros e baixo grau entre seus membros e os membros de grupos diferentes. Existem muitas formas de organizar n objetos em m grupos. Um dos problemas é devido ao fato de que m é geralmente desconhecido. Alguns métodos já foram pesquisados e para cada um deles vários algoritmos podem ser utilizados.

Os métodos são normalmente categorizados de acordo com o tipo de estrutura gerada, podendo ser hierárquicos ou não-hierárquicos.

Os métodos não-hierárquicos, ou partitivos, são mais simples, apenas dividindo os n objetos em m grupos sem qualquer sobreposição. Cada item, ao final do procedimento, pertence ao grupo que melhor representa suas características (FRAKES, 1992). Esses métodos são heurísticos por natureza, pois certas decisões precisam ser tomadas *a priori* como número de grupos, tamanho dos grupos e critérios de agrupamento. Isto fez com que não se utilizasse os métodos não-hierárquicos no presente estudo.

3.3.1 MÉTODOS HIERÁRQUICOS

Os métodos hierárquicos, por sua vez, são mais complexos, produzindo um conjunto de dados aninhados, em que pares de itens são sucessivamente ligados até que todos os itens do conjunto estejam conectados. A flexibilidade desses métodos consiste no uso de uma estrutura que armazene o co-relacionamento entre os grupos, permitindo ao usuário identificar grupos mais específicos ou abrangentes, conforme suas necessidades (WIVES, 2002). Os métodos hierárquicos podem ser:

- aglomerativos: com $n-1$ junções de pares inicialmente não agrupados (em grupos distintos) para formar um único grupo;
- divisivos: começando com todos os objetos em um mesmo grupo e $n-1$ divisões progressivas, resultando em n grupos com um objeto apenas.

Os métodos aglomerativos são amplamente utilizados em sistemas de recuperação de informações. (WANNER, 2004) descreve o seguinte algoritmo geral para os métodos de agrupamento aglomerativo:

a) iniciar cada objeto como um único grupo;

b) calcular a matriz de similaridades conforme descrito anteriormente;

c) unir os dois grupos mais similares (C_i e C_j) em um único grupo (C_{ij});

d) atualizar a matriz de similaridades, de forma a refletir as similaridades entre o novo grupo e os grupos restantes;

e) repetir os passos c e d até restar um único grupo;

A estrutura resultante do processo de agrupamento descrito é freqüentemente apresentada como um **dendrograma** (FIG. 5.2). Este representa os agrupamentos realizados, respeitando a ordem de união dos objetos e o nível de similaridade em que cada uma ocorreu (JAIN, 1999).

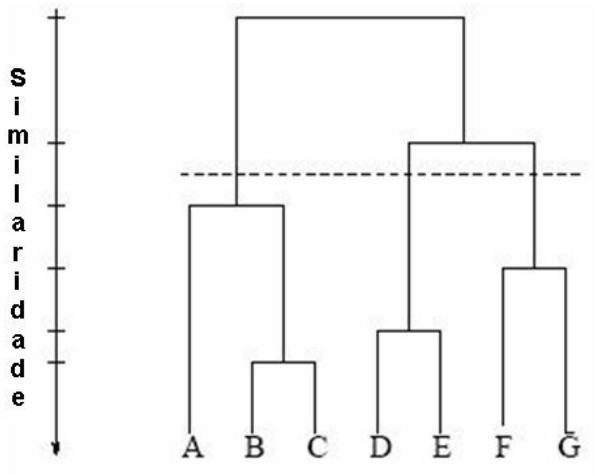

FIG. 3.3: Exemplo de um dendrograma

Dessa forma, o agrupamento unifica todos os documentos em um único grupo. É necessário então, determinar os grupos obtidos após este processo. O dendrograma indica os caminhos que a identificação de grupos deve seguir. É possível determinar um nível de similaridade para separação dos grupos ou simplesmente informar a quantidade desejada de grupos.

Os métodos aglomerativos se mostraram como uma boa solução para separar os criptogramas de acordo com a chave usada na cifragem.

No passo de atualização da matriz, novas similaridades $S_{ij,k}$, entre o grupo C_{ij} recém formado com todos os grupos C_k restantes, devem ser calculadas. Uma variedade de métodos para atualização de uma matriz foi desenvolvida e os métodos mais populares são **ligação simples** (*single link*), **ligação completa** (*complete link*) e **ligação por média dos grupos** (*group average link*) (CHANGE, 1998).

3.3.1.1 LIGAÇÃO SIMPLES

A escolha do método de agrupamento a ser utilizado, assim como de qualquer outro método de um modelo de RI, é feita empiricamente, objetivando obter os melhores resultados para uma determinada aplicação. Cada método possui suas particularidades que, dependendo da aplicação, podem até comprometer seus resultados.

O algoritmo de ligação simples gera bons resultados em grupos concêntricos (FIG. 3.4). Entretanto, em virtude do efeito cadeia, forma grupos não condizentes com a realidade em ambientes esféricos (FIG. 3.5) (JAIN, 1999).

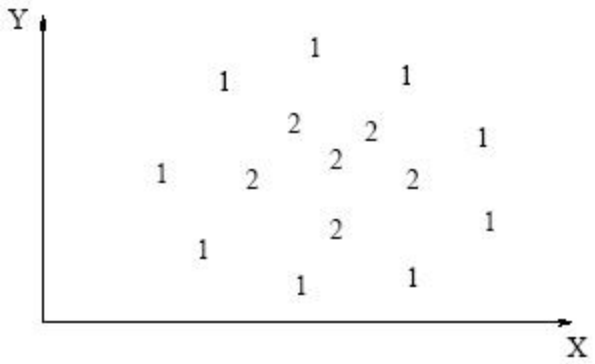

FIG. 3.4: Dois grupos concêntricos

No método de ligação simples, uma nova similaridade $S_{ij,k}$ é determinada pela maior das similaridades ($S_{i,k}$ e $S_{j,k}$) entre os grupos recém unidos (C_i e C_j) com o grupo C_k.

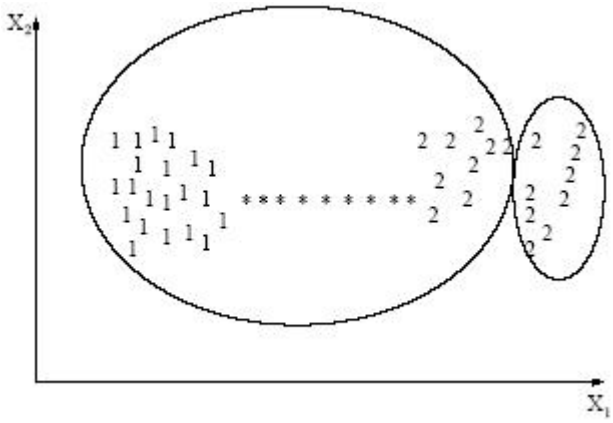

FIG. 3.5: Efeito do método de ligação simples em um ambiente esférico

Assim:

$$S_{ij,k} = \max(S_{i,k}, S_{j,k})$$

O dendrograma da FIG. 3.6 apresenta o resultado do agrupamento para o exemplo utilizado durante este capítulo (TAB. 3.3).

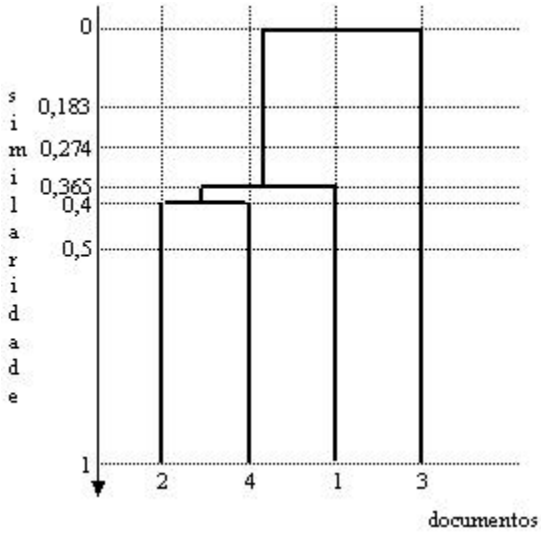

FIG. 3.6: Resultado do método de ligação simples

Este exemplo comprova o efeito de dispersão desse método. Observe que é fácil um usuário de um sistema de RI esperar que a escolha de um nível de similaridade, para a separação em grupos, garanta que todas as similaridades entre os documentos de um grupo sejam maiores que esse nível. Entretanto, no método de ligação simples, as similaridades entre textos de um mesmo grupo nem sempre são maiores que o nível escolhido. Por

exemplo, se o nível de similaridade escolhido for 0,3, os documentos um e dois pertenceriam ao mesmo grupo, apesar de a similaridade entre dois ser de apenas 0,183.

3.3.1.2 LIGAÇÃO COMPLETA

O método de ligação completa, ao contrário da ligação simples, produz grupos mais compactos, sendo mais eficiente em ambientes esféricos (FIG. 3.7) (JAIN, 1999).

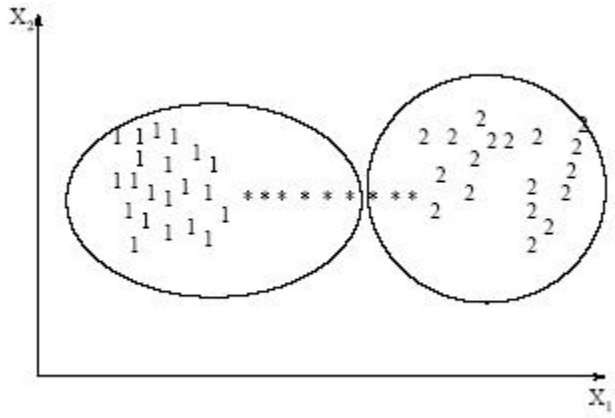

FIG. 3.7: Efeito do método de ligação completa em um ambiente esférico

No método de ligação completa, uma nova similaridade $S_{ij,k}$ é determinada pela menor das similaridades ($S_{i,k}$ e $S_{j,k}$) entre os grupos recém unidos (C_i e C_j) com o grupo C_k. Assim:

$$S_{ij,k} = \min(S_{i,k}, S_{j,k})$$

O dendrograma da FIG. 3.8 mostra o resultado do agrupamento, desse método, para o exemplo usado até aqui. Note que, ao contrário da ligação simples, o nível de similaridade escolhido garante um valor mínimo para a similaridade entre os documentos de um grupo.

Por outro lado, a similaridade entre documentos de grupos distintos pode ser maior que o valor determinado. Caso o nível de similaridade seja 0,3, os documentos um e quatro, apesar da similaridade ser de 0,365, pertenceriam a grupos distintos.

3.3.1.3 LIGAÇÃO POR MÉDIA DOS GRUPOS

No método de ligação por média dos grupos, por sua vez, todos os objetos contribuem para a nova similaridade, resultando em um efeito intermediário entre a amarração dispersa da ligação simples e a amarração compacta da ligação completa (FRAKES, 1992).

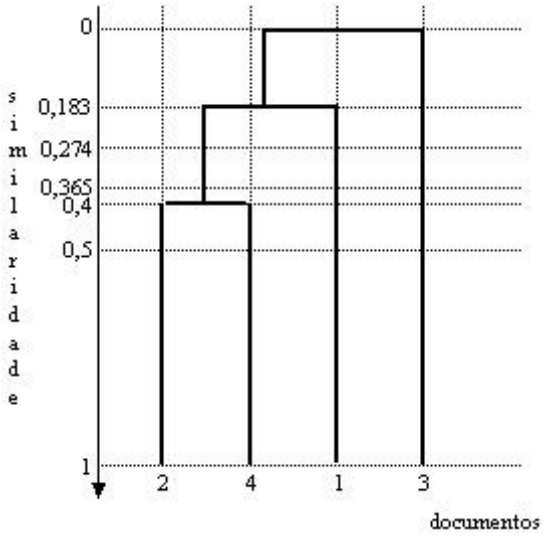

FIG. 3.8: Resultado do método de ligação completa

Uma nova similaridade $S_{ij,k}$ é determinada pela média ponderada das similaridades ($S_{i,k}$ e $S_{j,k}$) entre os grupos recém unidos (C_i e C_j) com o grupo C_k. Assim:

$$S_{ij,k} = \frac{m_i * S_{i,k} + m_j * S_{j,k}}{m_i + m_j},$$

em que m_i e m_j são os números de elementos dos grupos C_i e C_j respectivamente.

O resultado desse método é apresentado no dendrograma da FIG. 3.9.

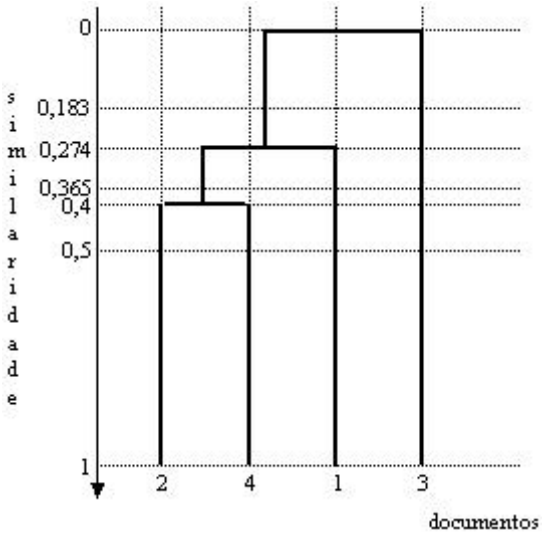

FIG. 3.9: Resultado do método de ligação por média dos grupos

3.4 CLASSIFICAÇÃO

As técnicas de classificação utilizam uma abordagem diferente para separar um conjunto de textos (ou ainda objetos) em grupos. Em contraste com o agrupamento, o objetivo é construir um mapeamento entre objetos de uma entrada em zero ou mais classes de um conjunto pré-definido de classes.

Um classificador aprende um padrão de classificação por meio de um treinamento prévio, normalmente com objetos não usados na classificação, identificando a(s) classe(s) desses objetos (LANGIE, 2003). O aprendizado ocorre de forma supervisionada, ou seja, os objetos são rotulados com sua(s) classe(s), geralmente por um processo manual, e submetidos ao classificador para que o mesmo identifique as características das classes existentes.

O raciocínio baseado em memória (*memory based reasoning*) é uma das técnicas existentes para a realização do aprendizado. A filosofia dessa técnica é a busca de uma solução para a situação atual baseada na recuperação de uma experiência passada, determinada pelo treinamento, semelhante. O k-NN, descrito a seguir, é um classificador que utiliza a estratégia do raciocínio baseado em memória para determinar a(s) classe(s) de um objeto.

Nesse contexto, para separar criptogramas conforme a chave de cifragem, seria necessário um treinamento em que os criptogramas fossem rotulados de acordo com a chave. A dificuldade de realizar um treinamento, com todo o espaço de chaves, inviabiliza o uso da classificação nesse processo.

Entretanto, em cifras polialfabéticas, é fácil observar que, quanto maior é o período da chave, maior é o vocabulário e menores são as similaridades entre os documentos. Essa relação permite afirmar que cada grupo, originado pelo agrupamento, possui criptogramas com o mesmo padrão de similaridade e tamanho de vocabulário. Assim, é possível identificar o tamanho da chave de um grupo por meio de técnicas de classificação.

Um estudo comparativo entre diversos algoritmos de classificação foi realizado em (YANG, 1999). O método que gerou melhores resultados foi o SVM (*Support Vector Machine*) (JOACHIMS, 1998). Entretanto, o SVM é usado quando existem apenas duas classes. Assim, a algoritmo considerado mais adequado para este trabalho foi o k-NN (*k-Nearest Neighbor*) (MITCHELL, 1996), com o qual também foram obtidos bons resultados.

3.4.1 *K-NEAREST NEIGHBOR*

O princípio de funcionamento do k-NN é a classificação de um objeto de acordo com os k objetos treinados mais próximos (vizinhos). Desta forma, é necessário utilizar uma métrica que represente adequadamente esta proximidade. A métrica deve ser definida pelo o sistema de RI respeitando as necessidades da aplicação.

Na situação aqui descrita, uma métrica possível é o valor absoluto da diferença entre as médias das matrizes de similaridades de dois grupos. Assim, os k vizinhos mais próximos de um grupo são aqueles em que as distâncias mais se aproximam de zero. Alguns experimentos são apresentados no Capítulo 4 procurando obter um procedimento mais eficiente para a determinação do período da chave.

O valor k pode ser definido pela aplicação ou pelo próprio usuário. É importante destacar apenas que o k deve ser (DUDA, 2000):

- grande o suficiente para minimizar a probabilidade de erro de classificação;
- pequeno o suficiente de forma a dar uma estimativa acurada da verdadeira classificação do novo objeto.

O k-NN pertence a uma classe de classificadores que não necessitam de uma função explícita de aprendizagem. O treinamento é feito, simplesmente, armazenando uma forma representativa de cada objeto de uma base de treino juntamente com sua(s) categoria(s). Assim, o treinamento do procedimento para determinação do período da chave pode ser o armazenamento do par <média das similaridades, tamanho da chave>. Um diferencial de um classificador pode ser a inclusão de cada classificação correta no conjunto de treinamento, caracterizando uma forma de aprendizado constante do classificador.

Observe que o k-NN recupera os k vizinhos mais próximos de um novo objeto. Assim, o classificador deve definir a política de determinação da(s) categoria(s) desse objeto. Nos experimentos utilizando o k-NN, os resultados foram expressos por duas políticas distintas:

- Unicategorização: o classificador indica como resultado um único tamanho de chave para um determinado grupo;
- Multicategorização: o classificador fornece vários tamanhos de chave possíveis para certo grupo.

A eficiência da classificação pode ser medida comparando o resultado fornecido com o de uma classificação manual. Um conjunto de teste é normalmente utilizado para avaliar a qualidade da classificação.

3.5 MEDIDAS DE AVALIAÇÃO

Os resultados de classificação, assim como os de agrupamento, necessitam ser avaliados. Uma variedade de experimentos foi realizada e seus resultados foram analisados segundo as recomendações de Yang e Liu (YANG, 1999).

A avaliação numérica dos resultados é medida pela abrangência (*recall*), precisão (*precision*) e F_1. A abrangência indica a relação entre a quantidade de elementos recuperados corretamente pelo sistema (cs) e a quantidade esperada de elementos recuperados (c). A precisão é a proporção entre a quantidade de elementos recuperados corretamente pelo sistema (cs) e o total de elementos recuperados pelo sistema (na). F_1 combina abrangência (r) e precisão (p) de modo que:

$$F_1(r,p) = \frac{2 * r * p}{r + p}$$

Essas medidas podem ser calculadas em macro-média (*macro-averaging*) ou micro-média (*micro-averaging*). No primeiro modo, a abrangência e a precisão são calculadas localmente (por grupo ou categoria) e só então a média é computada, fornecendo o resultado final. Assim, considerando n grupos (ou categorias) existentes, os valores de abrangência e precisão são dados por:

$$r = \frac{1}{n}\sum_{i=1}^{n}(cs_i/c_i)\,,\ p = \frac{1}{n}\sum_{i=1}^{n}(cs_i/na_i)\,,$$

em que cs_i é o número de atribuições corretas do sistema para o grupo i, c_i é o número real de elementos existentes no grupo i e na_i é o número de atribuições do sistema para o grupo i.

Em micro-média, os valores globais são calculados em vez dos relativos a cada categoria, assim:

$$cs = \sum_{i=1}^{m} cs_i\,,\ c = \sum_{i=1}^{m} c_i\,,\ na = \sum_{i=1}^{m} na_i$$

e

$$r = \frac{cs}{c}\,,\ p = \frac{cs}{na}$$

4 RI EM CRIPTOANÁLISE POLIALFABÉTICA

A conjectura de que os padrões lingüísticos sejam mantidos mesmo após a cifragem de um texto permite o uso de sistemas de RI em criptoanálise. A chave deve ser considerada como uma propriedade lingüística que determina o vocabulário da nova "linguagem".

Em termos de criptoanálise polialfabética, conforme estudo na Seção 2.2.2, o procedimento de determinação do período da chave converte o problema em várias cifras monoalfabéticas. Alguns procedimentos para a determinação desse período foram desenvolvidos. Entretanto a eficiência desses procedimentos diminui quando o tamanho da chave aumenta em relação ao tamanho do texto.

Assim, apesar de atualmente as cifras polialfabéticas estarem em desuso, é válido ainda o desenvolvimento de um processo novo de criptoanálise. Principalmente, em virtude de gerar conhecimento para a aplicação de novas tecnologias nas cifras contemporâneas.

A evolução dos experimentos realizados expõe aprimoramentos que culminam num processo que determina o tamanho da chave de forma mais eficiente que os métodos clássicos. Os procedimentos aqui formulados são baseados nas técnicas de agrupamento e classificação estudadas.

4.1 TREINAMENTO PARA DETERMINAÇÃO DO PERÍODO DA CHAVE

O método aqui proposto é baseado na repetição de padrões lingüísticos no decorrer dos criptogramas. Os padrões analisados são as palavras ou termos dos textos cifrados. A repetição de padrões em criptogramas distintos fornece indícios de que estes textos foram cifrados com uma mesma chave.

A FIG. 4.1 apresenta o fluxograma, em modo geral, para o treinamento e validação do procedimento de descoberta do período da chave.

Uma primeira preocupação é com relação aos textos em claro a serem utilizados. Todos os experimentos foram conduzidos sobre textos extraídos da Bíblia, por sua fácil disponibilidade na Internet e por conter uma vasta quantidade de textos sobre mesmo assunto. A realidade da atividade do criptoanalista permite que ele limite o escopo de assuntos possíveis, dado que normalmente a origem do criptograma é conhecida. Por exemplo, em situações militares, pode-se esperar a interceptação de mensagens solicitando

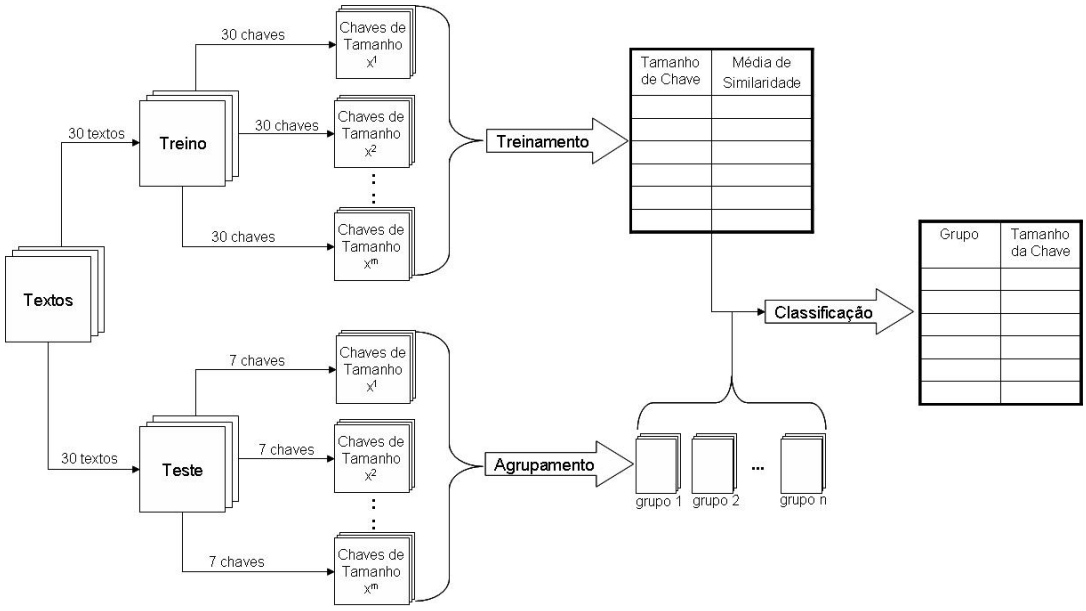

FIG. 4.1: Fluxograma do treinamento e validação do procedimento de determinação do período da chave

recursos (como comida, armamento e soldados).

As mensagens originais são de língua inglesa. Cada documento (em claro) foi extraído de capítulos da Bíblia (BIBLE, 2005) de forma a obter cerca de 2400 palavras por documento. As similaridades entre eles ficaram em torno de 0,87. Um total de 60 textos foi selecionado e dividido em duas partes iguais para gerar as duas coleções: a de treino e a de teste.

Cada coleção é gerada cifrando os textos com diversas chaves de diversos tamanhos. Para cada tamanho, 30 chaves pseudo-aleatórias foram geradas para cifrar a coleção de treino e sete chaves para a coleção de teste. As letras que compõem uma chave foram obtidas por meio de uma função de aleatorização disponível na linguagem Java. Os tamanhos de chaves utilizados variaram de acordo com o experimento realizado.

Observe que os textos em claro que geram a coleção de treino são diferentes dos que geram a coleção de teste. No processo de cifragem, é importante salientar que:

- os espaços foram preservados;
- as letras foram tratadas como minúsculas;
- qualquer símbolo, que não letra, foi ignorado.

Assim, os termos analisados são apenas palavras comuns. Entretanto, caso os espaços não sejam mantidos na cifragem, é possível estipular um tamanho fixo para os termos.

No treinamento do classificador k-NN (vide Seção 3.4), usado nos procedimentos descritos, é construída uma tabela contendo a média das similaridades entre os criptogramas cada grupo do conjunto de treino, rotulada com o tamanho da chave desse grupo. Um grupo é composto pelos textos que foram cifrados com uma mesma chave.

Encerrado o treinamento a validação é realizada por meio do agrupamento e classificação dos criptogramas da coleção de teste. O agrupamento tem como objetivo separar essa coleção em grupos de acordo a chave de cifragem. O agrupamento é feito conforme visto na Seção 3.3.1.

A determinação do período da chave é feita pela classificação de cada grupo obtido no agrupamento, respeitando o treinamento realizado. Assim, o resultado determina que todos os documentos de um grupo foram cifrados com uma mesma chave do tamanho encontrado.

Um conjunto de experimentos validou dois procedimentos distintos de determinação do período da chave. No primeiro procedimento, a classificação foi feita com o k-NN. Os resultados não foram satisfatórios e foi integrado o Índice Kasiski (Seção 4.3.1) ao segundo procedimento. Em um último experimento foi realizada uma comparação consistente do procedimento proposto com o método Kasiski, comprovando uma melhor eficiência do método aqui desenvolvido..

4.2 1º EXPERIMENTO

Um primeiro experimento foi realizado utilizando um classificador k-NN. Nesse experimento, os textos foram cifrados com chaves de tamanho 3, 5, 7, 9, 11, 13, 15, 17 e 19.

A TAB. 4.1 mostra uma análise preliminar da coleção de teste. Nesta, cada grupo de texto cifrado com uma mesma chave é examinado com respeito ao tamanho de vocabulário, ou seja, número de palavras de um criptograma, e média das similaridades. Em virtude do uso de várias chaves distintas de um mesmo tamanho, a média dos valores obtidos nas chaves de mesmo tamanho é calculada para a composição da tabela.

Uma chave longa implica que existiram diversas representações para uma mesma palavra em claro, dependendo de sua posição na mensagem. Este fato, comprovado pela análise da TAB. 4.1, justifica que o aumento no tamanho da chave implica no aumento do vocabulário e na diminuição das similaridades, comprovando a possibilidade de classificar

TAB. 4.1: Análise preliminar: vocabulário e similaridades

Período da Chave	Média do Vocabulário	Média de Similaridades
3	672	0,82939
5	852	0,7982
7	980	0,77263
9	1079	0,74722
11	1156	0,72365
13	1224	0,70314
15	1276	0,68555
17	1326	0,6691
19	1370	0,64715

um grupo de textos cifrados com uma mesma chave de acordo com o tamanho da mesma.

Nesse experimento, cada grupo obtido no agrupamento, representado também pela média das similaridades, é classificado para $k = 30$. Este valor é devido às 30 chaves utilizadas para cada tamanho. No resultado, de acordo com a política de unicategorização (Seção 3.4), o tamanho de uma chave é determinado pela categoria mais comum entre os k vizinhos. Por outro lado, na multicategorização, são retornados todos os tamanhos de chave encontrados entre os vizinhos.

4.2.1 RESULTADOS

Primeiramente, é necessário avaliar o resultado do agrupamento. Os métodos de ligação simples, completa e por média dos grupos foram aplicados. Apesar de os grupos serem definidos a partir da entrada de um nível de similaridade ou a quantidade desejada de grupos, a entrada foi simplesmente a quantidade real de grupos existentes. O nível de similaridade não foi usado devido à variação entre os valores de similaridades, proporcionada pela diversidade de tamanhos de chaves.

O agrupamento produziu grupos perfeitos. Assim, cada grupo foi composto por todos os documentos cifrados com a mesma chave e em nenhum grupo ocorreram textos cifrados com chaves distintas. A divergência entre os vocabulários gerados por chaves distintas permite o sucesso do agrupamento, com precisão e abrangência iguais a um. Um raciocínio análogo seria pensar em agrupar textos de línguas diferentes, que resultariam em grupos com textos apenas de uma mesma língua.

O resultado da classificação é apresentado na TAB. 4.2, composta pelos valores da

avaliação feita sobre as políticas de uni- e multicategorização. Como esperado, a precisão na multicategorização é baixa, devido à variedade de tamanhos de chaves dos vizinhos de um grupo. Entretanto, para o criptoanalista pode ser interessante possuir essa diversidade de tamanhos, compensada pela melhora da abrangência do sistema.

TAB. 4.2: Resultado da classificação do experimento 1

	miP	miR	miF_1	maP	maR	maF_1
Unicategorização	0,95238	0,95238	0,95238	0,95679	0,95238	0,95458
Multicategorização	0,69663	0,98413	0,81579	0,75351	0,98413	0,85351

miP = precisão micro-media
miR = abrangência micro-media
miF_1 = F_1 micro-media
maP = precisão macro-media
maR = abrangência macro-media
maF_1 = F_1 macro-media

Em uma análise por tamanho de chave (TAB. 4.3), observa-se que os resultados tendem a piorar com o aumento do tamanho da chave.

TAB. 4.3: Análise da classificação dos resultados por categoria

Tamanho da Chave	3	5	7	9	11	13	15	17	19
Abrangência	1	1	1	1	1	1	1	0,7143	0,8571
Precisão	1	1	1	1	1	1	0,7778	0,8333	1

Neste contexto, aparentemente os resultados podem ser considerados satisfatórios. Contudo, este procedimento não é mais eficiente que o método Kasiski, que calculou corretamente o período da chave em todos os documentos da coleção de teste. Na execução do método Kasiski foram analisadas as seqüências de dez letras.

4.3 2º EXPERIMENTO

Algumas modificações no treinamento do classificador foram realizadas na busca por um procedimento mais eficiente, inclusive para chaves de tamanhos maiores. Assim, treinamentos com base no tamanho do vocabulário, cálculos sobre a matriz de similaridades e até mesmo uma combinação desses elementos foram testados. Entretanto, em nenhum caso, a eficiência teve uma melhora considerável.

Em virtude do classificador *k-NN* não gerar os resultados desejados, uma etapa intermediária, baseada no método Kasiski, foi incorporada ao procedimento. Dessa forma, para

cada grupo obtido no agrupamento é calculado o Índice Kasiski (IK), conforme descrito a seguir, que restringe os vizinhos possíveis do mesmo.

4.3.1 ÍNDICE KASISKI

O método Kasiski, conforme Seção 2.2.2.1, indica uma única possibilidade para o tamanho de uma chave. Todavia, é razoável supor que o criptoanalista disponha de alguns tamanhos de chaves distintos para testar. O Índice Kasiski é definido em DEF. 5.1.

DEF. 5.1: No cálculo do Índice Kasiski, assim como no método Kasiski, são calculados todos os fatores das diferenças entre seqüências idênticas de um criptograma. O mais freqüente fator primo é definido como Índice Kasiski. O maior fator primo é usado como critério de desempate, quando vários fatores são os mais freqüentes.

O método Kasiski, assim como o IK, é executado sobre um único documento. Assim para calcular o Índice Kasiski de um grupo (IK_c) é necessário calcular o índice de cada documento (IK_d) desse grupo. Posteriormente, o IK_c é determinado pelo valor de IK_d mais freqüente dentro do grupo. Neste momento, é esperado que o período da chave de todos os documentos de um grupo seja múltiplo de IK_c.

Na classificação, o treinamento do k-NN permanece inalterado, conforme o primeiro experimento. No teste, então, os vizinhos mais próximos de um grupo a ser classificado são escolhidos entre aqueles em que o tamanho da chave é múltiplo do IK_c encontrado.

4.3.2 RESULTADOS

Neste experimento, foi utilizada uma maior variedade de tamanhos de chave. Os documentos foram cifrados com todas as chaves de tamanho ímpares variando de três a 99.

O agrupamento da coleção de teste foi realizado e, conforme esperado, gerou grupos perfeitos, compostos por todos os documentos cifrados com a mesma chave. Entretanto, os resultados desejados foram obtidos apenas com o uso do método de ligação completa.

Pôde-se observar ainda, que os valores de IK_c calculados, para cada grupo resultante do agrupamento, sempre revelaram um fator correto para o tamanho da chave. Entretanto, em algumas situações, o IK_d calculado não pertencia ao conjunto de fatores do tamanho da chave, justificando, assim, o ganho de eficiência quando se trabalha com uma coleção de criptogramas.

A TAB. 4.4 mostra, então, os valores da avaliação feita sobre as políticas de uni- e multicategorização, descritas anteriormente.

TAB. 4.4: Resultado da classificação do experimento 2

	miP	miR	miF_1	maP	maR	maF_1
Unicategorização	0,92420	0,92420	0,92420	0,94184	0,92420	0,93294
Multicategorização	0,90582	0,95335	0,92898	0,93952	0,95335	0,94638

miP = precisão *micro-averaging*
miR = abrangência *micro-averaging*
miF_1 = F_1 *micro-averaging*
maP = precisão *macro-averaging*
maR = abrangência *macro-averaging*
maF_1 = F_1 *macro-averaging*

A medida que melhor avalia o resultado aqui exposto é a de abrangência, por comparar o resultado do sistema com o resultado real esperado. Os resultados de abrangência, em valores absolutos, foram piores do que o experimento anterior. Todavia, este procedimento é mais eficiente por trabalhar com chaves de tamanhos maiores. Os resultados obtidos foram semelhantes aos do método Kasiski, em que a abrangência foi de 0,98299.

4.4 3º EXPERIMENTO

Uma análise dos resultados obtidos nos experimentos anteriores indica a necessidade de uma comparação mais consistente com o método Kasiski. Apenas por meio dessa comparação, pode ser comprovado que o procedimento proposto é mais eficiente que o Kasiski.

Dessa forma, o procedimento executado no experimento anterior deve ser repetido para chaves de tamanhos maiores em relação ao tamanho do texto. Os textos utilizados nos outros experimentos são grandes, cerca de 2400 palavras. Assim, optou-se por reduzir o tamanho dos textos para cerca de 1000 **letras**. No experimento, foram utilizadas chaves de tamanhos ímpares variando de três a 49. O período da chave possui, então, uma relação de até 5% do tamanho do texto.

A TAB. 4.5 mostra uma análise preliminar da coleção de teste semelhante a apresentada na TAB. 4.1. Neste caso, foram examinados ainda valores com respeito a mediana das similaridades de um grupo cifrado com uma mesma chave. Pode então ser observada a mesma tendência de comportamento do primeiro experimento, em que com o aumento do período a chave, os valores das similaridades diminuem e o vocabulário aumenta.

TAB. 4.5: Análise preliminar da coleção do experimento 3

Período da Chave	Média das Similaridades	Mediana das Similaridades	Média do Vocabulário
3	0,5278	0,5268	181,9667
9	0,2973	0,2955	228,3952
15	0,2102	0,2098	242,6048
21	0,1672	0,1652	250,9571
27	0,139	0,137	255,381
33	0,1207	0,1183	258,3667
39	0,1092	0,1082	161,5429
45	0,099	0,0983	263,3143

4.4.1 RESULTADOS

Assim como no experimento anterior, o agrupamento produziu resultados perfeitos de precisão e abrangência quando utilizado o método de ligação completa. Com o agrupamento, foram produzidos 168 grupos de textos cifrados com uma mesma chave.

Na fase de rotulação, o IK_d revelou um fator correto do tamanho da chave para períodos menores que 19. Com o aumento do tamanho da chave, começam a aparecer erros no cálculo de IK_d. Esses erros, todavia, são minimizados com o calculo de IK_c. Apenas em 4 dos 168 grupos gerados o IK_c obtido não correspondeu a um fator do tamanho da chave. Esses erros ocorreram em chaves de tamanho 47. Nesse experimento, foram investigadas seqüências de três letras para o cálculo do IK.

Nesse experimento, para a determinação dos vizinhos mais próximos, foram usadas, além da média das similaridades, as seguintes métricas: mediana, distância de euclides entre matrizes de similaridades, vocabulário e 2-norma.

As TABs. 4.6 e 4.7 apresentam os resultados sobre as políticas de uni- e multicategorização respectivamente.

Os resultados obtidos, exceto quando a 2-norma é utilizado, são satisfatórios. A distância de Euclides apresentou-se como a melhor solução em termos de unicategorização. Na multicategorização, por outro lado, os resultados com o vocabulário foram ligeiramente melhores.

Uma comparação entre a abrangência do procedimento proposto e o método Kasiski é essencial para comprovar que o método aqui desenvolvido é mais eficiente que o Kasiski. A TAB. 4.8 mostra a comparação dos resultados de unicategorização com o método Kasiski separados por tamanho de chave.

TAB. 4.6: Resultado de unicategorização no experimento 3

	miP	miR	miF_1	maP	maR	maF_1
Média	0,9329	0,9107	0,9217	0,9477	0,9107	0,9288
Mediana	0,9024	0,881	0,8917	0,9282	0,881	0,904
Euclides	0,9939	0,9702	0,9819	0,9948	0,9702	0,9824
Vocabulário	0,9756	0,9524	0,9639	0,9823	0,9524	0,9671
2-Norma	0,5732	0,5595	0,5663	0,8472	0,5595	0,674

miP = precisão *micro-averaging*
miR = abrangência *micro-averaging*
miF_1 = F_1 *micro-averaging*
maP = precisão *macro-averaging*
maR = abrangência *macro-averaging*
maF_1 = F_1 *macro-averaging*

TAB. 4.7: Resultado de multicategorização no experimento 3

	miP	miR	miF_1	maP	maR	maF_1
Média	0,8723	0,9762	0,9318	0,9142	0,9762	0,9442
Mediana	0,8454	0,9762	0,9162	0,897	0,9762	0,9349
Euclides	0,8962	0,9762	0,9452	0,9352	0,9762	0,9552
Vocabulário	0,9011	0,9762	0,948	0,9432	0,9762	0,9594
2-Norma	0,3249	0,7619	0,4588	0,6088	0,7619	0,6768

A análise dessa tabela é suficiente para comprovar que o procedimento proposto é bem mais eficiente que o método Kasiski. Observa-se ainda que os resultados do Kasiski não são satisfatórios em chaves de tamanho não primos. Entretanto, os períodos de chave encontrados nesses casos são normalmente fatores primos do tamanho real da chave, justificando assim nossa definição de Índice Kasiski.

É importante ressaltar que, com o aumento do tamanho da chave em relação ao tamanho do texto, torna-se mais difícil o trabalho da criptoanálise. Aqui não é diferente, até porque, em virtude do Índice Kasiski, é necessária a repetição de seqüências de caracteres no decorrer de um criptograma para o cálculo do IK. Entretanto, obteve-se um ganho significativo de eficiência no procedimento aqui desenvolvido.

O tempo de execução é um outro fator que pode ser levado em consideração durante os experimentos. Entretanto, inclusive pela escolha por métodos hierárquicos de agrupamento que são mais complexos, o tempo não foi uma preocupação no presente estudo, sendo necessária apenas a execução dos experimentos em um tempo considerado satisfatório. Em geral, os experimentos não levaram mais do que algumas horas para serem executados, sendo o processamento dos textos a etapa mais custosa de todo o processo.

TAB. 4.8: Valores de abrangência para todos os tamanhos de chave

Período da Chave	Euclides	Vocabulário	Média	Kasiski
3	1	1	1	1
5	1	1	1	1
7	1	1	1	1
9	1	1	1	0,0095
11	1	1	1	1
13	1	1	1	1
15	1	1	1	0
17	1	1	1	1
19	1	1	1	0,9667
21	1	1	1	0
23	1	1	1	0,9857
25	1	1	1	0,0048
27	1	1	1	0
29	1	1	1	0,8667
31	1	1	1	0,9333
33	1	1	1	0
35	1	0,8571	0,8571	0
37	1	1	1	0,7857
39	1	0,5714	0,2857	0
41	1	1	1	0,7286
43	1	1	1	0,7085
45	0,8571	1	0,2857	0
47	0,4286	0,4286	0,4286	0,5
49	1	1	1	0,0238
Média	0,9702	0,9524	0,9107	0,5214

5 RI EM CIFRAS DE BLOCOS

Outro enfoque deste estudo foram os sistemas contemporâneos de criptografia, em especial as cifras de bloco. Ao tratar sobre cifras de blocos, estará implícito que o modo de operação utilizado é o ECB, que foi usado durante os experimentos aqui conduzidos. Neste, uma mensagem é dividida em blocos, e cada bloco é cifrado separadamente de acordo com a chave e o algoritmo utilizado. O criptograma é obtido concatenando os blocos cifrados.

O princípio fundamental das cifras de bloco é a provável eliminação da correlação entre os dados de entrada (chave e texto em claro) com os de saída (criptograma). Uma cifra é considerada confiável desde que atenda a um conjunto de requisitos definidos pela comunidade. Entretanto, os testes propostos são basicamente de reprovação.

5.1 AGRUPANDO CRIPTOGRAMAS DE ACORDO COM A CHAVE

A não existência de forma de avaliação, que garanta a segurança de uma cifra, motiva a busca por falhas em algoritmos bem aceitos pela comunidade.

Em uma cifra de blocos, se um mesmo bloco é cifrado duas vezes, com a mesma chave, os dois blocos resultantes serão iguais. Esta repetição permite a identificação de criptogramas cifrados com uma mesma chave. Assim, com a realização do agrupamento (FIG. 5.1), sobre uma coleção de criptogramas, é esperada a obtenção grupos apenas com documentos que foram cifrados com uma mesma chave, mesmo sem o conhecimento da mesma.

O alfabeto utilizado pelas cifras de bloco é o binário, em que os espaços entre palavras do texto em claro não são reconhecidos, impossibilitando a segmentação correta do texto em palavras, como nas cifras clássicas. Assim, o termo, unidade textual elementar para os sistemas de RI, é um bloco de tamanho predefinido. Este tamanho depende do algoritmo usado e vale, por exemplo, 64 para o DES e 128 para o AES.

Em cifras clássicas, o espaço de termos existente é bastante limitado, constituído apenas por seqüências, normalmente pequenas, de letras. É então comum haver termos iguais em textos cifrados com chaves distintas. Em cifras de blocos, essa situação dificilmente ocorre, devido à diversidade de termos possíveis de serem gerados. É importante relem-

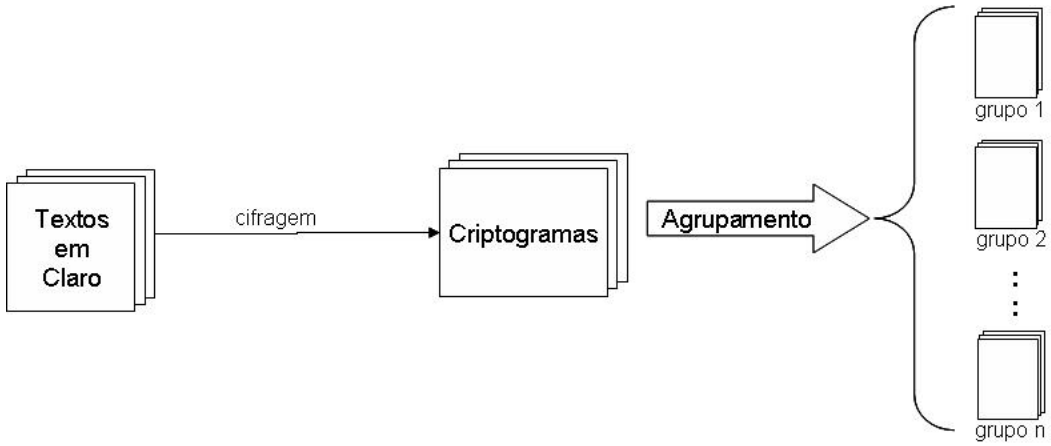

FIG. 5.1: Fluxograma para realização do agrupamento

brar que o conjunto de termos possíveis em cifras de blocos é da ordem de 2^m, onde m é o tamanho do bloco em bits. Assim, caso a similaridade entre dois criptogramas quaisquer seja maior que zero, significando que há pelo menos um bloco comum aos mesmos, então os mesmos possivelmente foram cifrados com uma mesma chave.

Em contrapartida, eventualmente ocorre que a similaridade seja zero, mesmo entre dois textos cifrados com a mesma chave. Torna-se necessária então a aplicação de uma técnica de agrupamento que permita a existência de documentos em que a similaridade entre eles seja zero em um mesmo grupo. O método de ligação simples, conforme pôde ser observado na Seção 3.3.1.1, respeita esta política. Assim, é possível formar grupos em que existam valores de similaridades, entre textos de um mesmo grupo, abaixo de um parâmetro de corte. Neste contexto, apenas o método de ligação simples foi utilizado no decorrer dos experimentos.

5.2 AVALIAÇÃO EXPERIMENTAL

Um conjunto de três experimentos, usando cifras de bloco tradicionais como o DES e o AES, foi realizado para verificar o efeito do agrupamento sob algumas abordagens. Um último experimento foi ainda feito com criptogramas disponibilizados por uma terceira entidade, comprovando assim a força do processo proposto. Os três primeiros experimentos utilizam um conjunto de criptogramas derivado de textos da Bíblia (BIBLE, 2005).

Os conceitos de precisão e abrangência (vide Seção 3.5) foram utilizados para avaliar os grupos gerados. A precisão mede a proporção entre as atribuições corretas feitas pelo

sistema e o total de atribuições feitas. Esta medida é usada para verificar os grupos formados contêm apenas textos cifrados com uma mesma chave.

No contexto das cifras de blocos, como dito anteriormente, é bastante difícil a existência de termos iguais cifrados com chaves distintas. Assim, dificilmente textos cifrados com chaves diferentes pertencerão ao mesmo grupo. Com isso, a precisão esperada é normalmente um, caracterizando um resultado excelente no processo de agrupamento.

Por outro lado, a abrangência é a proporção entre as atribuições corretas feitas pelo sistema dividido e o número esperado de atribuições corretas. Esta medida é utilizada para determinar o quanto se conseguiu agrupar do total de textos cifrados com uma mesma chave. Observe que as medidas de precisão e abrangência se completam, objetivando recuperar grupos apenas com textos cifrados com a mesma chave e grupos com a maior quantidade possível de textos cifrados com a mesma chave.

Nos experimentos aqui realizados, mesmo conhecendo a quantidade de grupos realmente existentes, a definição dos grupos no processo de agrupamento teve que ser feita por meio de um parâmetro de corte, ou seja, o nível de similaridade. A FIG. 5.2 apresenta um resultado possível no presente contexto.

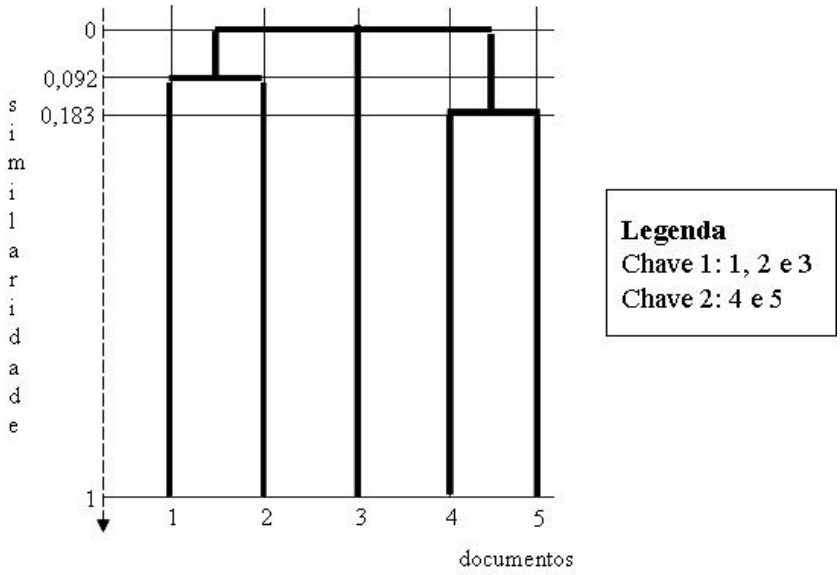

FIG. 5.2: Exemplo possível de resultado do agrupamento

Mesmo sabendo que apenas duas chaves distintas foram utilizadas, o dendrograma apresenta claramente três grupos, em que a similaridade entre documentos de grupos diferentes é sempre zero. Assim, informando que são desejados apenas dois grupos, dois

dos três grupos realmente existentes serão unidos de uma forma desconhecida, podendo então misturar textos cifrados com chaves distintas.

Vista a inexatidão dos resultados quando a quantidade de grupos é informada, optou-se como parâmetro de corte o valor imediatamente maior que zero. Assim, nos experimentos realizados, nem sempre o agrupamento retornou apenas um único grupo com textos cifrados com a mesma chave. Isto implica, logicamente, numa redução da abrangência.

Com a existência de vários grupos representando a mesma chave, optou-se por calcular a abrangência, de uma determinada chave, apenas sobre o grupo composto pela maior quantidade de criptogramas. Contudo, a precisão continua sendo calculada normalmente sob todos os grupos gerados.

5.2.1 1º EXPERIMENTO

Um primeiro experimento foi realizado para avaliar a influência do tamanho das mensagens e dos textos originais na eficiência do agrupamento. Nesse experimento, em cada etapa (FIG. 5.3) foram extraídos da Bíblia 30 textos de tamanhos iguais. Estes textos foram cifrados, com o DES ou o AES, usando 50 chaves pseudo-aleatórias. Assim como nos experimentos do capítulo anterior, as chaves foram geradas por meio de uma função de aleatorização disponível na linguagem Java. O agrupamento é feito, então, sobre um total de 1500 criptogramas, objetivando separá-los por chave.

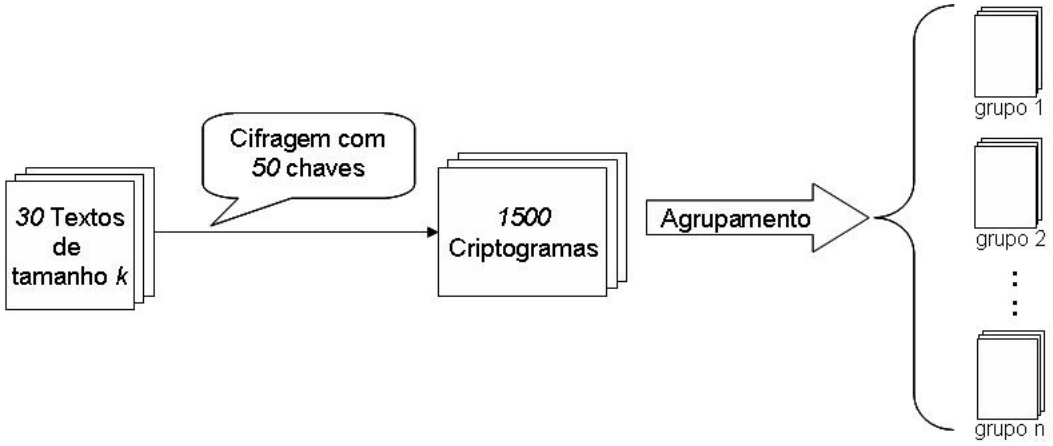

FIG. 5.3: Fluxograma para execução do 1º experimento

Esse processo é executado diversas vezes, para vários tamanhos de textos diferentes. Assim, é possível verificar um tamanho recomendável para que o agrupamento seja rea-

lizado com sucesso. A TAB. 5.1 apresenta os resultados, de precisão e abrangência, com textos cifrados pelo DES.

TAB. 5.1: Influência do tamanho do texto na eficiência do agrupamento - DES

Tamanho dos Textos (bytes)	Precisão	Abrangência
10240	1	1
8192	1	1
6144	1	1
4096	1	1
2048	1	1
1024	1	1
512	1	1
256	1	0.8
192	1	0.71
128	1	0.4
64	1	0.6

Como esperado, os valores de precisão foram sempre iguais a um, indicando que todos os grupos contêm apenas documentos cifrados com uma mesma chave. Os resultados só foram perfeitos, entretanto, em criptogramas de pelo menos 512 bytes. Nesses tamanhos, conforme confirmado pelos valores de abrangência, todos os textos cifrados com uma mesma chave foram colocados no mesmo grupo.

Em tamanhos de menores que 512 bytes, textos cifrados com uma mesma chave, em alguns momentos, pertenceram a grupos distintos. Os valores de abrangênca também indicam uma tendência de diminuição da eficiência com a diminuição dos tamanhos dos criptogramas. Esses resultados comprovam a influência do tamanho da mensagem no sucesso do processo de agrupamento.

A TAB. 5.1 não apresenta, entretanto, os resultados de cada chave em separado. Para cada tamanho, os textos utilizados na cifragem de cada uma das chaves foram iguais. Como efeito, os grupos resultantes do agrupamento foram rigorosamente idênticos, conforme o esboço de uma situação possível (FIG. 5.4). Isto indica que um grupo, representando uma chave qualquer, possui grupos equivalentes, de outras chaves, compostos exatamente pelos mesmos documentos originais. Esse resultado comprova a influência do texto em claro no processo de agrupamento.

Os resultados com os testes realizados com o AES são mostrados na TAB. 5.2. O AES é uma cifra de bloco de 128 bits, enquanto o DES é de apenas 64 bits. Essa diferença no

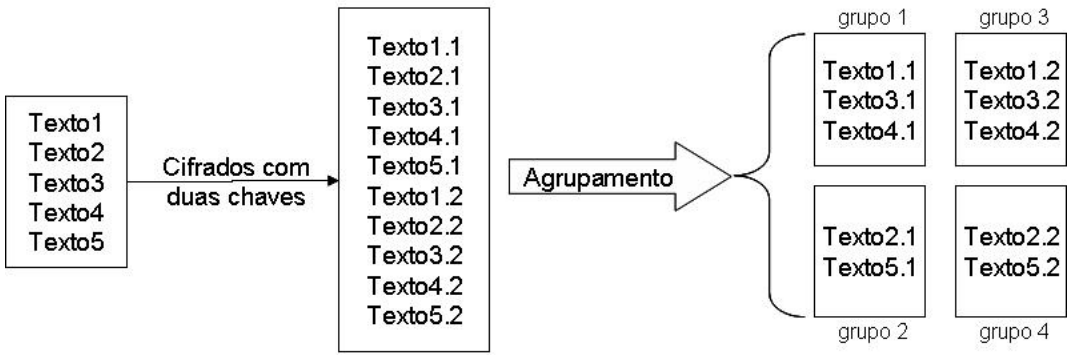

FIG. 5.4: Resultado possível de agrupamento nas condições do 1º experimento

tamanho do bloco implicou em um aumento considerável no tamanho dos criptogramas para que o agrupamento fosse realizado com sucesso.

TAB. 5.2: Influência do tamanho do texto na eficiência do agrupamento - AES

Tamanho dos Textos (bytes)	Precisão	Abrangência
10240	1	1
8192	1	1
6144	1	1
4096	1	0.97
3072	1	0.87
2560	1	0.5
2048	1	0.33
1536	1	0.2
1024	1	0.16

A identificação dos tamanhos dos criptogramas em bytes diminui um pouco a noção de tamanho dos mesmos. Imagine, então, que para o agrupamento do DES são necessários textos que não ocupam sequer meia página, enquanto para o agrupamento do AES são necessárias quase duas páginas de texto.

A necessidade, um tanto lógica, de textos maiores para a realização com sucesso do agrupamento no AES é em virtude de o espaço de termos ser bem maior no AES do que no DES. Conforme visto na Seção 2.3, enquanto no AES tem-se 2^{128} termos possíveis, o DES tem apenas 2^{64} termos possíveis.

Neste experimento, é notória uma fraqueza das cifras de blocos. Os resultados obtidos com o agrupamento comprovam a existência de uma correlação entre a chave e o texto em claro com o criptograma gerado em uma cifra de bloco. Mesmo utilizando o AES,

os tamanhos dos textos, necessários para um agrupamento com sucesso, são plenamente viáveis.

5.2.2 2º EXPERIMENTO

As cifras de bloco são, normalmente, caracterizadas por executar a cifragem de uma mensagem em várias iterações. A cada rodada, os bits da chave e da mensagem são misturados de forma que, ao final do processo, cada bit do criptograma dependa de todos os bits da mensagem e da chave.

Por outro lado, os projetos de cifras, como o DES, não justificam o porquê do uso de 16 rodadas, e não 13 ou 20 rodadas, por exemplo. Assim, considerou-se necessária a avaliação do processo de agrupamento com relação à quantidade de rodadas de uma cifra.

O experimento anterior foi repetido, utilizando os mesmos textos e a mesmas chaves. O DES foi, então, modificado para executar com duas, cinco, oito, 11, 20 ou 32 rodadas. Essa variação não provocou impacto nos resultados, que permaneceram como a TAB. 5.1.

Em virtude da "pouca" mistura dos bits da mensagem com o da chave, talvez fosse esperado, que o resultado do agrupamento fosse menos preciso, quando uma quantidade pequena de interações fosse realizada. Contudo, mesmo com poucas rodadas o agrupamento já consegue identificar correlações entre o criptograma e o texto claro e a chave utilizada.

Nos métodos tradicionais de criptoanálise, é comum que o trabalho da criptoanálise seja mais eficiente quando uma quantidade menor de rodadas é realizada pela cifra. Um efeito negativo, é que o resultado da abrangência não melhorou com a diminuição do número de iterações. Entretanto, em nenhum momento, dois blocos distintos cifrados com uma mesma chave podem gerar um mesmo bloco cifrado, e esta seria a única forma que permitiria aumentar a abrangência.

5.2.3 3º EXPERIMENTO

Os experimentos realizados até aqui possuem algumas particularidades para verificar as potencialidades e limitações do processo de agrupamento. Agora, é necessária também a realização do agrupamento em uma situação mais próxima da real, conforme o fluxograma da FIG. 5.5.

Este experimento introduziu uma maior variabilidade na coleção de teste: 300 textos de dez tamanhos distintos foram cifrados com o DES utilizando 20 chaves pseudo-aleatórias

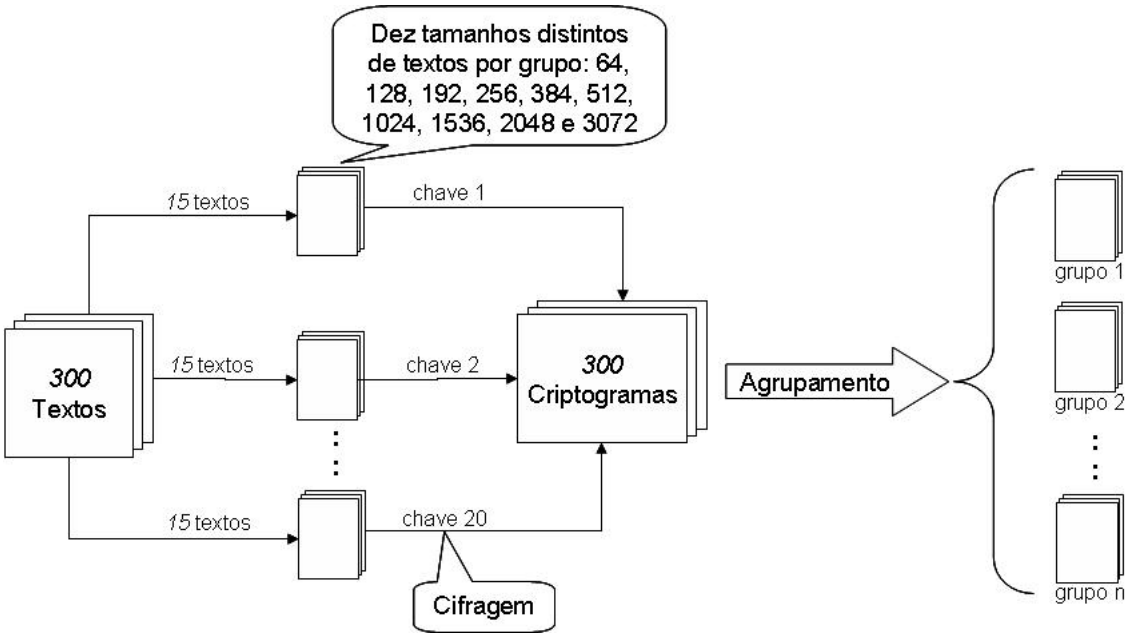

FIG. 5.5: Fluxograma para execução do 3º experimento

(15 textos para cada chave). Os tamanhos, em bytes, das mensagens foram 64, 128, 192, 256, 384, 512, 1024, 1536, 2048 e 3072.

Observe que tamanhos, em que o agrupamento pode não ser realizado com sucesso, foram usados. Se apenas tamanhos "adequados" fossem utilizados, a abrangência esperada do agrupamento seria sempre um. Assim é possível analisar o comportamento dos grupos resultantes do agrupamento, verificando se os criptogramas maiores conseguem corrigir os erros de abrangência provocados pela existência de textos menores.

Os resultados mostraram que a precisão continuou inalterada e os valores de abrangência são apresentados na TAB. 5.3.

TAB. 5.3: Abrangência em criptogramas de tamanhos variados

Abrangência
0.66 em 3 das chaves
0.8 em 2 das chaves
0.87 em 3 das chaves
0.93 em 10 das chaves
1 em 2 das chaves

Os resultados aqui expostos comprovam que, usando criptogramas de tamanho inadequados, nem sempre é possível obter a abrangência desejada. Entretanto, em dois casos, os grupos obtidos estavam completos (abrangência igual a um), e em outros dez, apenas

um único criptograma ficou isolado do seu grupo. Assim, nesse contexto os resultados podem ser considerados satisfatórios.

5.2.4 4º EXPERIMENTO

Um último experimento foi realizado para simular um trabalho de interceptação de criptogramas. Com a colaboração de uma terceira entidade, pôde-se verificar a correção do agrupamento, quando não se tem conhecimento das mensagens originais, chaves e até mesmo da cifra utilizada na cifragem.

Em virtude principalmente da limitação em relação ao tamanho das mensagens, algumas restrições foram estabelecidas. Foi solicitado, um conjunto de 200 criptogramas resultante da cifragem de 20 textos em claro (de 10 kbytes cada) com dez chaves distintas. O bloco, de 64 bits, era o único conhecimento que se tinha sobre o algoritmo de criptografia.

O agrupamento foi realizado conforme proposto na FIG. 5.1. Um outro agrupamento foi realizado nas mesmas condições acima, mas foi usada uma cifra de 128 bits. Em ambos os casos, os resultados foram como previsto pela primeira experiência, com precisão e abrangência iguais a um.

Este experimento comprova uma grande vantagem do processo de agrupamento, que pode ser estendido, em princípio, a qualquer cifra de bloco.

No contexto atual, as cifras de blocos são projetadas de forma que a mensagem e chave sejam suficientemente misturadas, gerando um criptograma "aleatório", sem qualquer correlação com os dados de entrada. O sucesso na separação de criptogramas de acordo com a chave provoca um forte questionamento sobre essa aleatoriedade, identificando uma correlação entre os dados de entrada e saída. Assim, o uso de uma cifra bem aceita na comunidade não garante a segurança desejada. É importante levar em consideração também modo de operação utilizado na cifragem. É fundamental que este modo ofereça algum nível de segurança, o que não ocorre quando o modo ECB é aplicado.

Em testes preliminares realizados com uma cifra operada em modo CBC (Cipher Block Chaining), a separação de criptogramas de acordo com a chave não foi possível. Isso se deve ao fato que a saída da cifragem de um bloco é usada na cifragem do bloco seguinte, eliminando a repetições dos blocos. Desse modo, os resultados comprovam que o modo CBC é mais seguro que o ECB.

É importante salientar que, mesmo usando o modo CBC, o resultado da precisão

foi positivo. Os agrupamentos realizados, conforme esperado, sempre geraram grupos somente com textos cifrados com uma mesma chave. Assim, pode ser possível encontrar alguma correlação entre os dados de entrada e saída de uma cifra operada em modo CBC.

O tempo de execução, assim como nos experimentos realizados no Capítulo 4, não foi uma preocupação no presente estudo. Entretanto, é importante que o trabalho de criptoanálise seja executado em tempo hábil. Os experimentos foram realizados em, no máximo, poucas horas, o que é perfeitamente satisfatório.

6 CONSIDERAÇÕES FINAIS

As técnicas de recuperação de informações podem ser aplicadas em diversos contextos. Este estudo apresentou o uso inovador das mesmas em criptoanálise. Um procedimento foi desenvolvido para determinação do período da chave em cifras polialfabéticas. Além disso, foi comprovada uma fraqueza das cifras de blocos operadas em modo ECB.

Em cifras polialfabéticas, alguns procedimentos foram desenvolvidos até a obtenção de um que determinasse de forma mais eficiente o período da chave. O grande diferencial do procedimento criado é a aplicação de técnicas de agrupamento, que permitem a separação de criptogramas de acordo com a chave usada. O cálculo do Índice Kasiski foi ainda integrado a um classificador k-NN para possibilitar a determinação do tamanho da chave. Uma outra vantagem desse procedimento é que as etapas a eles pertinentes são independente da linguagem.

Nos métodos de determinação do período da chave, a eficiência dos mesmos tende a diminuir com o aumento do tamanho da chave. Contudo, nos experimentos realizados, o procedimento proposto apresentou-se bem mais eficiente que o método Kasiski.

Uma cifra de bloco se propõe a gerar criptogramas com distribuição aleatória de símbolos, ou seja, com base somente nos criptogramas, não deveria ser possível obter algo a respeito da chave. Entretanto, devido a repetições de blocos, é possível determinar quais textos foram cifrados com uma mesma chave.

A determinação desses textos foi aqui feita por técnicas de agrupamento. A separação de textos de acordo com a chave usada na cifragem confirmou a hipótese que chaves diferentes determinam linguagens distintas para seus criptogramas.

Alguns experimentos foram realizados avaliando os resultados do agrupamento quanto à precisão e à abrangência. Assim, foi possível determinar o tamanho de textos necessário para que um agrupamento seja realizado com sucesso. Provou-se ainda que o agrupamento é imune a variação de rodadas de uma cifra e que pode ser aplicado mesmo quando o algoritmo de cifragem é desconhecido.

A aplicação de técnicas de RI em cifras de bloco é a maior contribuição dessa dissertação. Por meio dessas técnicas foi provado, experimentalmente, que os criptogramas gerados por cifras de blocos, operadas em modo ECB, não são realmente aleatórios. A

realização com sucesso do agrupamento mostra a existência de uma correlação entre os dados de entrada com os de saída de uma cifra.

Como conseqüência, é extremamente necessário que, na escolha por um sistema criptográfico, seja também avaliado o modo de operação a ser usado. Se a segurança do modo ECB já era questionada, seu uso agora é altamente não-recomendável.

Duas aplicações evidentes das técnicas aqui empregadas são a identificação de emissor e a de detecção de mudanças de chave. Novas investigações ainda podem ser realizadas, na tentativa de se extrair outras informações a respeito da chave.

Uma outra possibilidade é a utilização em ataques com texto em claro conhecido. Pode-se, por exemplo, cifrar um corpus com todas as chaves possíveis e cada novo criptograma capturado pode ser classificado em um dos conjuntos de chaves existentes, identificando, assim, qual foi a chave usada. Uma redução do espaço de chaves a ser utilizado pode ser obtida por meio dos métodos tradicionais de criptoanálise das cifras de blocos. A integração com esses métodos evita o trabalho exaustivo da força bruta, sendo, então, mais eficiente.

Alguns experimentos com o modo de operação CBC foram realizados, comprovando que o mesmo é mais seguro que o ECB. Contudo, não foi eliminada a possibilidade de que seja identificada uma correlação entre dados de entrada e saída.

7 REFERÊNCIAS BIBLIOGRÁFICAS

BIBLE in basic english. Disponível: http://www.o-bible.com/bbe.html [capturado em dezembro de 2005].

BIHAM, E.; BIRYUKOV, A. **An improvement of Davies' attack on DES**. Journal of Cryptology, v.10, n.3, p.195–205, 1997.

BIHAM, E.; SHAMIR, A. **Differential cryptanalysis of the Data Encryption Standard**. New York: Springer-Verlag, 1993.

BIHAM, E.; SHAMIR, A. **Differential cryptanalysis of des-like cryptosystems**. Journal of Cryptology, v.4, n.1, p.3–72, 1991.

BIRYUKOV, A. **Block ciphers and stream ciphers**: the state of the art. 2004.

CHANG, C.-H.; HSU, C.-C. **Hypertext information retrieval for short queries**. In Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop, 1998.

DAVIES, D. W. **Investigation of a potential weakness in the DES algorithm**. Private Communications, 1987.

DENNING, D. E. R. **Cryptography and data security**. Addison-Wesley Publishing Company, USA, 1982.

DUDA, R. O. ET. AL. **Pattern classification**. Wiley-Interscience, 2000.

FEISTEL, H. **Cryptography and computer privacy**. Scientific American, v. 228, n.5, p. 15–23, 1973.

FIPS 46-3 - Data Encryption Standard (DES). NIST, Outubro de 1999.

FIPS 81 - DES modes of operation. NIST, Dezembro de 1980.

FIPS 197 - Advanced Encryption Standard (AES). NIST, Novembro de 2001.

FRAKES, W. B.; YATES, R. B. **Information retrieval**: data structures and algorithms. Prentice Hall, 1992.

GASPERIN, C. V. **Fundamentos do processamento estatístico da linguagem natural**. Trabalho Individual, PUC-RS, 2000.

GEBBIE, S. **A Survey of the mathematics of cryptology**. Tese de Mestrado, University of the Witwatersrand, South Africa, Fevereiro de 2002.

JAIN, A. K. ET. AL. **Data clustering**: a review. ACM Computing Surveys, v. 31, n. 3, p. 264–323, 1999.

JAIN, A. K.; DUBES R. C. **Algorithms for clustering data**. Prentice Hall, 1988.

JAKOBSEN, T.; KNUDSEN L. R. **The interpolation attack on block ciphers**. In Fast Software Encryption, LNCS 1267, Springer-Verlag, p. 28–40, 1997.

JOACHIMS, T. **Text categorization with support vector machines**: learning with many relevant features. In European Conference on Machine Learning (ECML), Springer-Verlag, p. 137–142, 1998.

KAHN, D. **The codebreakers**: the story of secret writing. Macmillan Publisshing Co. Inc., New York, 1967.

KRAAIJ, W.; POHLMANN R. **Viewing stemming as recall enhancement**. In Proceedings of SIGIR'96, p. 40–48, 1996.

LAMBERT, J. A. **Cifrador simétrico de blocos**: projeto e avaliação. Tese de Mestrado, IME, 2004.

LANGFORD, S. K.; HELLMAN, M. E. **Differential-linear cryptanalysis**. In Advances in Cryptology: Proceedings of CRYPTO'94, Springer-Verlag, p. 17–25, 1994.

LANGIE, L. C.; LIMA, V. L. S. **Classificação hierárquica de documentos textuais digitais usando o algoritmo kNN**. In 1º Workshop em Tecnologia da Informação e da Linguagem Humana, v. 1, p. 1–10, São Carlos, 2003.

LINK, S. **Information systems security**. Block Course, Dept of Information Systems, Massey University, 2003.

MANNING, C.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge, MA: MIT Press, 1999.

MATSUI, M. **Linear cryptanalysis method for DES cipher**. In Advances in Cryptology: Proceedings of EUROCRYPT'93, Springer-Verlag, p. 386–397, 1993.

MENEZES, A. J. ET. AL. **Handbook of applied cryptography**. CRC Press, 1996.

MITCHELL, T. **Machine learning**. McGraw Hill, 1996.

MITKOV, R. **The Oxford handbook of computational linguistics**. New York: Oxford University Press, 2005.

MORAES, R. **Construção de um ambiente Web com ferramentas para estudo de algoritmos de criptografia através do MATLAB**. UFRJ, 2004.

PORTUGUÊS (do brasil): freqüência de ocorrência das letras. versão 1.2, 2002. Disponível: http://www.numaboa.com.br/criptologia/matematica/estatistica/freqPortBrasil.php [capturado em dezembro de 2005].

RUBIN, A. D. **An experience teaching a graduate course in cryptography**. Cryptologia, v. 21, n. 2, p. 97-109, 1997.

RUKHIN, A. ET. AL. **A statistical test suite for the validation of cryptographic random number generators**. NIST Computer Security Division/Statistical Engineering Division Internal Document, Setembro de 1999.

SALTON, G.; BUCKLEY, C. **Term weighting approaches in automatic text retrieval**. Technical Report, Cornell University, 1987.

SALTON, G. **Introduction to modern information retrieval**. McGraw-Hill, 1983.

SCHNEIER, B. **Applied cryptography**. John Wiley & Sons, 1996.

SINGH, S. **O livro dos códigos**. Record, 2001.

SPILLMAN, R. ET. AL. **Use of genetic algorithms in cryptanalysis of simple substitution ciphers**. Cryptologia, v. 17, n. 1, p. 31–44, 1993.

STANDAERT, F.-X. ET. AL. **Cryptanalysis of block ciphers**: a survey. Technical Report, Université Catholique de Louvain, 2003.

VAN RIJSBERGEN, C. J. **Information retrieval**. Dept. of Computer Science, University of Glasgow, 1979.

WAGNER, D. **Towards a unifying view of block cipher cryptanalysis**. In Fast Software Encryption 2004, invited paper, 2004.

WANNER, L. **Introduction to clustering techniques**. 2004.

WIVES, L. K. **Técnicas de descoberta de conhecimento em textos aplicadas à inteligência competitiva**. Exame de Qualificação, UFRGS, 2002.

WIVES, L. K. **Um estudo sobre técnicas de recuperação de informações com ênfase em informações textuais**. Trabalho Individual, UFRGS, 1997.

YANG, Y.; LIU, X. **A re-examination of text categorization methods**. In Proceedings of SIGIR'99, p. 42–49, 1999.

YANG, Y.; PEDERSEN, J. O. **A comparative study on feature selection in text categorization**. In Proceedings of ICML'97, p. 412–420, 1997.

YATES, R. B.; NETO, B. R. **Modern information retrieval**. New York: Addison Wesley, 1999.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo