CODAGE et PROTECTION DE L'INFORMATION

Dans ce chapitre, nous allons voir les méthodes classiques permettant de coder des informations par des mots écrits sur {0,1}*.

1 - NOTIONS DE CODAGE

Un *code* est une application entre l'ensemble des informations à représenter et un ensemble de configurations binaires.

Le *codage* est l'opération qui réalise la transformation des informations en séquences binaire.

Le *décodage* est l'opération qui transforme les séquences binaires en informations intelligibles. Un code est *homogène* si tous les mots de code ont la même longueur, c'est le cas des codes en téléinformatique.

L'ensemble des caractères à transmettre constitue un alphabet. La représentation binaire d'un caractère est un mot de code.

Propriétés souhaitables

Soient N la taille de l'alphabet et n la longueur des mots de code.

- 2^n doit être le plus proche possible de N
- n doit être aussi petit que possible
- Il doit être possible de protéger des erreurs de transmission
- Le décodage doit être aisé
- Les tris alphabétiques doivent être aisés.

2 - QUELQUES NOTIONS SUR LES CODES

Supposons que U soit un bloc de k bits à transmettre

$$U=u_1u_2 \ldots u_k$$

et que le codeur transforme U en C un bloc de n bits

$$C=c_1c_2 \ldots c_n \ .$$

Un tel code est appelé *code (n,k)* .

Le code est *systématique* si pour $1 \leq i \leq k$, on a $c_i = u_i$. Les bits $c_{k+1} \ldots c_n$ sont appelés bits de contrôle. Le code est *en bloc* si le mot C ne dépend que de U. Dans le cas contraire, on dit qu'il est *convolutionnel* .

Le *rendement* du code est la quantité

$$R = \frac{k}{n} \ .$$

Le *poids de Hamming* d'un mot de code est le nombre de 1 qu'il contient.

La *distance de Hamming* entre deux mots de code est le poids du vecteur somme (le nombre de bits en lesquels ils diffèrent).

Codes linéaires

Un code (n,k) est *linéaire* s'il est systématique et si les n-k bits de contrôle dépendent linéairement des k bits d'information.

Il existe une matrice H dans $\{0,1\}^{(n-k)\times n}$ appelée *matrice de contrôle* telle que

$$H\,C^T = 0 \quad \text{avec } H=(A, I_{n-k})$$

où A est une matrice (n-k)xk et I_{n-k} est la matrice identité (n-k) x (n-k).

On appelle *matrice génératrice* d'un code la matrice G telle que

$$C=U\,G \text{ avec } G = (I_k , -A^T).$$

Exemples

$$H1 = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$H2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$H_3 = (1\ 1\ 1\ 1)$$

Théorème. *Un code linéaire de distance minimum entre les mots* d *permet de corriger un nombre d'erreurs*

$$e = \left\lfloor \frac{d-1}{2} \right\rfloor$$

et dans le cas pair de détecter une erreur supplémentaire.

Exemples (suite) $e_1 = 1$, $e_2 = 2$, pour H_3 détection simple d'une erreur.

Si C est le mot transmis, le récepteur reçoit en réalité un mot R tel que R = C + E où E est *le vecteur d'erreur*. Le décodeur doit restituer avec une probabilité d'exactitude la plus forte possible le vecteur R+E. On appelle *syndrome* le produit $H\,R^T$.

Idée de l'algorithme

Si $H\,R^T = 0$ alors R est considéré comme un mot du code sinon une erreur s'est produite. On déclenche une alarme ou on procède à une correction directe en choisissant un vecteur d'erreur de poids minimal tel que $H\,E^T = H\,R^T$.

Les *codes de Hamming* sont des codes linéaires tels que toutes les colonnes de H sont non nulles et distinctes. Ces codes permettent la correction d'une erreur.

Codes cycliques

Un *code cyclique* est un code linéaire tel que toute permutation cyclique d'un mot de code est un mot du code.

Exemple .

$$H_4 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

Les mots d'un code cyclique sont représentés par un polynôme

$$c(x) = \sum_{i=0}^{n-1} c_{i+1}\, x^i \quad .$$

Propriété. *Tout polynôme représentant un mot d'un code cyclique* (n,k) *est un multiple d'un polynôme générateur* g(x) *de degré* n-k *associé à la dernière ligne de la matrice génératrice* G. *Ce polynôme divise* $x^n + 1$.

Pour le codage, on utilise la propriété suivante

Propriété. *Tout mot de code* c(x) *représentant une information* u(x) *est de la forme*

$$c(x) = x^{n-k}\, u(x) + r(x)$$

avec r(x) *le reste de la division de* x^{n-k} u(x) *par le polynôme générateur* g(x).

Pour le décodage, on utilise la propriété suivante

Propriété. *Pour un code cyclique de polynôme générateur* g(x), *le syndrome de E est le reste de la division du polynôme associé* e(x) *par* g(x).

Il est possible de calculer dans le cas de syndrome non nul un polynôme localisateur.

Codes polynomiaux

Un *code polynomial* est un code linéaire dont chacun des polynômes associés aux mots de code sont divisible par un polynôme générateur.

Exemple. L'avis V41 du CCITT conseille l'utilisation de codes polynomiaux de longueurs 260, 500, 980, 3860 bits avec le polynôme générateur

$$g(x) = x^{16} + x^{12} + x^5 + 1.$$

Un tel code permet de détecter 1 ou 2 erreurs, tous les paquets d'erreurs de longueur supérieure ou égale à 16 avec une probabilité supérieure à 99,99%.

C'est ce type de code qui est utilisé par les trames à tous les niveaux pour calculer le "checksum". Son intérêt est que le calcul peut être réalisé par des registres à décalages.

3 - CODAGE DES INFORMATIONS

Le code CCITT n° 5 est issu du code ASCII (7 bits) complété par 1 bit de parité ou d'imparité. C'est celui que l'on utilise couramment pour coder les caractères.

Certains codes permettent de compresser les données. Les codes de Huffman permettent de représenter les caractères avec des mots de code de taille inversement proportionnelle à la fréquence des caractères. Ils vérifient la propriété d'être préfixe, si u et v sont deux mots de code alors $u \neq vw$ et $v \neq uw$.

Si $X = \{x_1,...,x_n\}$ est l'alphabet avec $\forall\ 1 \leq i \leq n$, la lettre x_i a pour fréquence f_i. Soit $c(x_i)$ le code de x_i . Le code d'Huffman pour l'alphabet X doit minimiser la quantité

$$\lambda = \sum_{1 \leq i \leq n} f_i \, |c(x_i)|.$$

La construction se fait par un arbre binaire dont les feuilles sont étiquetées par les symboles à représenter et tels que la hauteur de la feuille est inversement proportionnelle à la fréquence de la lettre.

Pour les images, il existe de nombreux types de codage. Le codage Alphamosaïque est donné par la norme ISO 2022 (videotext). On découpe l'image en lignes et en colonnes et à chaque intersection on trouve un caractère. Ces caractères ont été normalisés on trouve

Le codage géométrique est celui qui représente l'image comme un ensemble de formes pour lesquelles on précise position et paramètres de taille.

Le codage numérique est une description point par point. Chaque point a un niveau de gris ou de couleur. Les plus classique sont RVB et TSL (voir figure 1).

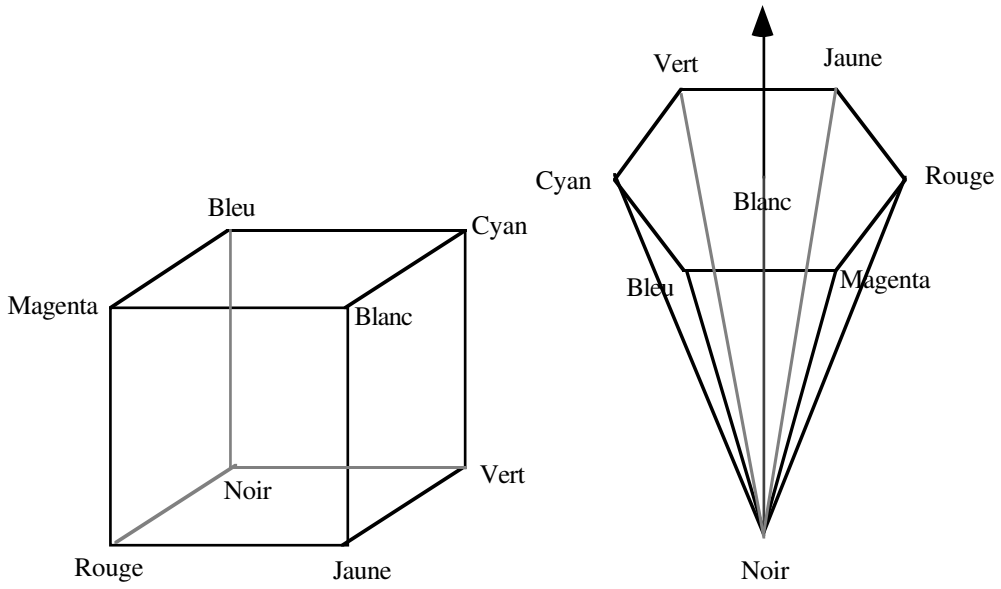

Figure 1. Codage RVB et TSL.

4 - PROTECTION DU CARACTERE PRIVE

La *garantique* est l'utilisation commerciale des techniques de sécurité.

La *cryptographie* est l'ensemble des méthodes utilisées pour garantir la *confidentialité* des informations. Le plus souvent les deux interlocuteurs disposent d'une clé. On distingue

- les systèmes *à clé privée* ,
- les systèmes *à clé publique* .

Ces systèmes travaillent soit en découpant le texte en petits morceaux et en y ajoutant des séquences de bits pseudo-aléatoires, soit sur tout le bloc d'information de telle sorte qu'une petite altération change totalement le message chiffré.

L'*authentification* est le second grand problème dans le domaine de la sécurité des données: chaque usager doit pouvoir produire des messages dont l'authenticité peut-être vérifiée par tous mais ne peuvent être produits que par lui-même. Elle se fait en général par l'utilisation de mots de passe (avec éventuellement plusieurs niveaux).

Systèmes à clé privée

L'émetteur chiffre son texte en utilisant un opérateur inversible S_K dépendant d'une information (appelée clé) K. La clé K est transmise par un canal privé au récepteur. Celui-ci peut déchiffrer le message reçu en utilisant l'opérateur inverse S_K^{-1}.

Ces systèmes utilisent des opérations de

- substitution, remplacement de n bits par n autres bits,
- transposition, permutation de groupes de n bits dans le texte en clair.

Systèmes à clé publique

Un système à clé publique est une paire d'algorithmes (E_K , D_K) tels que

- E_K est l'inverse de D_K ,
- pour toute valeur de K, E_K et D_K sont aisément calculables,
- il est impossible de calculer D_K à partir de E_K et D_K est gardé secret,
- connaissant K, il est impossible de trouver un couple (E_K ,D_K).

La clé K se trouve en général dans un annuaire public. Ces systèmes fournissent une réponse aux problèmes suivants:

- confidentialité,
- intégrité,
- authentification.

5 - QUELQUES SYSTEMES A CLES PRIVEES

Le système DES (Data Encryption Standard) utilise une clé de 56 bits. L'algorithme transforme un bloc de 64 bits en un bloc chiffré de 64 bits. Il se déroule en 19 étapes:

- une transposition indépendante de la clé sur les 64 bits,
- 16 itérations pendant lesquelles une clé de 48 bits est sélectionnée qui participe au chiffrage,
- échange des 16 bits de poids fort et de ceux de poids faible,
- une transposition inverse de la première exécutée.

Soit un couple (P,C) où C est le chiffre de P. Des algorithmes simples de calcul de la clé K échouent:

- calculer pour toute clé K le chiffrage de P demanderait un temps moyen en $O(2^{28})$!!
- pré-calculer pour toute clé K le chiffrage de P demanderait un stockage de 2^{56} enregistrement.

L'algorithme DES a été normalisé par l'ISO sous le nom DEA1 (Data Encipherment Algorithm number one).

Cependant lers spécialistes s'accordent à penser qu'une machine à déchiffrer ne vaut à l'heure actuelle que 1 million de dollars alors qu'elle en valait 4 en 1979.

Le triple DES consiste à utiliser trois passes de DESS avce deux clés différentes. Il est très utilisé dans les transactions bancaires.

RC2 et RC4 sont des algorithmes propriétaires de RSA Data Security Inc. Ils sont non publiés et donc pas évalués.

IDEA est un chiffrement développé à Zurich par Massey et Lai utilisant des clés à 128 bits de publication récente (1990). Sauf bouleversement des machines les spécialistes pensent qu'ils sont sur pour longtemps.

Skipjack est un algotithme secret à usage civil développé par la NASA. Il est sur pour une dizaine d'années. Sa spécificité est d'être implémenté sur des puces afin de permettre l'écoute légale de la transmission de données.

6 - QUELQUES SYSTEMES A CLES PUBLIQUES

Le système Merckle-Diffie-Hellman est basé sur un problème NP bien connu: la recherche dans un ensemble de nombre de sous ensembles de somme donnée.

La clé de codage A du destinataire est telle que

$$A=(a_1, a_2, ..., a_n)$$

avec pour tout i dans [1..n] le nombre a_i est entier. Pour la construire, le destinataire choisit deux entiers w et m premiers entre eux ainsi qu'une suite d'entiers

$$A'=(a'_1 , a'_2 , ..., a'_n) .$$

telle que pour i dans [2..n]

$$a'_i > \sum_{k=1}^{i-1} a'_k$$

alors il calcule

$$a_i = a'_i \ w \bmod m .$$

Exemple

n=10, w= 764, m=2731

A'=(3,5,11,20,41,83,169,340,679,1358),

A=(2292,1089,211,1625,1283,599,759,315,2597,2463).

L'émetteur découpe le message à émettre en bloc de n bits

$$X= (x_1, x_2, ..., x_n).$$

Le bloc d'information émis est alors

$$Y = AX = \sum_{i=1}^{n} a_i x_i$$

Pour décoder le message le destinataire calcule

$$Z = Y \ w^{-1} \bmod m.$$

L'algorithme consiste alors à décomposer Z sur la base de A' en tenant compte du fait que les a'_i sont en ordre croissants.

Algorithme

pour i allant de n à 1 faire

si $a'_i > Z$ alors début

$x_i := 1;$

$Z := Z - a'_i$

fin

sinon $x_i := 0;$

L'algorithme RSA du nom de ses inventeurs Rivest, Shamir et Adleman est l'un des sytèmes les plus puissants connus, il est utilisé par PGP.

Cet algorithme utilse cinq nombres :

- deux très grand nombre premier p et q ainsi que leur produit n,
- la clé de chiffrement e telle que e est premier avce (p-1)(q-1),
- la clé de chiffrement d telle que $d=e^{-1} \bmod ((p-1)(q-1))$.

Les nombres p et q restent secrets. L'algorithme chiffre toute information qui peut être représentée sous une forme numérique m inférieure à n.

La valeur chiffrée est

$$c = m^e \bmod n,$$

le déchiffrement s'opére en appliquant la formule

$m = c^d \bmod n$.

Ainsi la clé publique est (n,e).

La puissance de RSA repose sur la difficulté de factoriser n et l'absence de méthode algébrique permettant de calculer d à partir de données et de la clé publique.

Actuellement compte-tenu de la puissance des machines, les temps de factorisation sont les suivants :

Taille clé	temps sur machine 100MIPS
426 bits	14,5s
512 bits	22mn
700 bits	153 jours
1024 bits	280 000 ans