Teoria codurilor și criptografia
Codificarea informațiilor

Lorenzo Javier Martin Garcia

20 ianuarie 2022

1. Codificarea canalelor

Teoria informației studiază fluxul de informații de la un emițător la un receptor, printr-un canal.

Transmi ător Canal Receptor

Ca o caracteristică suplimentară, informațiile trebuie stocate și tratate în așa fel încât hârtiile receptorul și emițătorul pot fi schimbate dacă este necesar.

În general, informația care este transmisă provine dintr-o sursă care generează secvențe, s, de simboluri,

$$s = X_1X_2 \cdots X_n.$$

De exemplu, X_n poate fi al n-lea simbol al unui mesaj sau rezultatul celei de-a n-a iterații a unui experiment. În practică, această secvență va fi întotdeauna finită, dar în scopuri teoretice este uneori util să se ia în considerare secvențe infinite.

Să presupunem că fiecare simbol X_n este un element al unei mulțimi finite,

$$S = \{s_1,s_2,\dots,s_q\},$$

pe care îl vom numi alfabetul sursă.

Pentru simplitate, vom presupune că probabilitatea, p_i , ca simbolul al n-lea din succesiune să fie si ,

$$Pr(X_n = s_i) = p_i ,$$

rămâne fix în timp –este staționar– și depinde doar de i și nu de poziția pe care o ocupă în lanț –nu are memorie–. Astfel, diferite simboluri pot avea probabilități diferite de apariție, dar nu depind de simbolurile precedente din secvență. În orice caz, trebuie îndeplinit

$$i = 1,2,\dots,n \quad p_i \quad 0 \text{ și } \sum_{i=1}^{ce} p_i = 1.$$

În termeni statistici, S este o secvență de n variabile aleatoare independente și distribuite identic.

În general, simbolurile sursă în formatul lor original nu pot fi transmise pe canalul disponibil, pt care este necesar pentru a le codifica.

Pentru a codifica alfabetul sursă este folosit un set finit,

$$T = \{t_1,t_2,\dots,t_r\}$$

numit alfabet de cod și care conține simboluri de cod r. Evident, aceste simboluri depind de tehnologia canalului de transmisie.

Codurile care folosesc un alfabet de codare de simboluri r se numesc coduri r-ary, pentru De exemplu, dacă T = {0,1}, codul este binar.

Un cuvânt cod este o secvență finită de elemente ale alfabetului cod.

Codarea constă în asocierea unui cuvânt cod fiecărui simbol al alfabetului sursă, astfel încât să poată fi transmise de un anumit canal sau de alte circumstanțe.

Știind cum să codifice toate simbolurile sursă, un cuvânt sursă este codificat ca succesiune de cuvinte de cod asociate cu simbolurile sursă corespunzătoare care îl compun, fără separare între ele.

Un exemplu de codare este codul Morse unde alfabetul sursă este format din litere și cifre, iar alfabetul codului este format din punct, liniuță (un semnal de trei ori durata punctului) și absența unui semnal (cu durata unui punct între simboluri, trei puncte între litere și trei liniuțe între cuvinte) pentru a separa simbolurile, literele și cuvintele. Se știe că cuvântul SOS este codificat ca ··· — ···

Pentru a simplifica nomenclatura și atâta timp cât nu există confuzie, vom folosi „cuvânt” pentru a ne referi la cuvintele cod. Astfel, un cuvânt w este o secvență finită de simboluri a lui T și lungimea sa este numărul de simboluri ale alfabetului de cod care îl compun, $\ell = |w|$.

Mulțimea tuturor cuvintelor se notează cu T din , inclusiv cuvântul gol sau de lungime zero. A stabilit toate cuvintele diferite de zero se notează T +. da tu ^{n} = T ×T ×T ×T atunci

$$T = \bigcup_{n \quad 0} T^n \quad \text{Si t}^{+} = \bigcup_{n>0} T^n .$$

Un cod C este o funcție C : S 7 T + injectivcare asociază un cuvânt diferit de zero format din elemente ale lui T fiecărui simbol al lui S.

Injectivitatea asigură că fiecare simbol al lui T va fi codificat diferit de celelalte, ceea ce poate permite decodarea. Dacă această proprietate nu este îndeplinită, decodarea produce ambiguități și face imposibilă munca receptorului.

Multe proprietăți ale codurilor depind doar de cuvintele lor și nu de funcția de codificare în sine, astfel încât în aceste cazuri codul este considerat a fi imaginea funcției C.

Dacă S este definit asemanator cu T ,

$$\text{da} = \bigcup_{n \quad 0} \text{da}^n \quad \text{si } S^{+} = \bigcup_{n>0} \text{da}^n ,$$

funcția C poate fi extinsă la o altă funcție între S Si t ,

$$C \ : \text{da} \ 7 \quad T$$

astfel încât fiecare cuvânt sursă este codificat ca succesiune de cuvinte asociate cu simbolurile sale. În această schemă, deși teoretic nu trebuie să fie așa, din motive de coerență cuvântul sursă nul trebuie transformat în cuvântul cod nul. Imaginea acestei noi funcții este setul

$$\text{Img}(C \ ) = \{w i1 w i2 \text{ câștig} \quad T \quad : w ij \quad C \ n \quad 0\}.$$

Dacă lungimea cuvintelor wi asociate fiecărui simbol sursă q este notată cu ℓi , dacă , cel lungimea medie a codului C este definită ca

$$L(C) = \sum_{i=1}^{ce} p_i \ell_i .$$

Scopul teoriei codurilor este de a construi coduri

- a căror decodare este ușoară și lipsită de ambiguitate
- lungimea medie este cât se poate de mică.

Exemplul 1.1 De exemplu, în bazinele de fotbal, simbolurile S = {1,X,2} sunt folosite pentru a reprezenta victoria echipei gazdă (1), egalitatea (X) și victoria echipei în deplasare (2).

Dacă doriți să transmiteți aceste rezultate ale meciurilor de fotbal pe un canal binar, va trebui să codificați elementele lui S folosind elementele alfabetului T = {0,1}. , foarte simbolic ar putea fi

Un posibil cod, C′ : S 7 T

C (1) = 10, C (X) = 11, C (2) = 01.

Cele paisprezece rezultate ale unui grup, s = 1X112X112111X1, ar fi codificate ca cuvânt

w1 = 1011101001111010011010101110

de lungime ℓ1 = |w1| = 28.

Un alt cod posibil, C′′ : S 7 T , mai puțin simbolic ar putea fi

C „(1) = 0, C „(X) = 1, C „(2) = 00

unde rezultatele grupului anterior ar fi codificate ca cuvânt

w2 = 0100001000000010

de lungime ℓ2 = |w2| = 16.

Un al treilea cod, C′′′ : S 7 T ar putea fi

C "′(1) = 0, C "′(X) = 10, C "′(2) = 11

unde rezultatele grupului anterior ar fi codificate ca cuvânt

de lungime ℓ3 = |w3| = 19.

Dacă prin vreo procedură s-a stabilit că probabilitatea ca echipa gazdă să câștige un joc este p1 = 0,6, ca să remizeze este p2 = 0,25 și să piardă este p3 = 0,15, lungimile medii ale codurilor C′ , C și C′′′ ar fi

) = 2 0,6+2 0,25+2 0,15 = 2,

L(CL(C ") = 1 0,6+1 0,25+2 0,15 = 1,15, L(C "′)

= 1 0,6+2 0,25 +2 0,15 = 1,4.

Fiecare dintre aceste coduri are caracteristici diferite care le fac adecvate sau inadecvate pentru un anumit scop.

2. Decodați firele în mod unic

Un cod C este decodabil în mod unic dacă și numai dacă fiecare element al lui T corespunde lui 7 T este sub C cel mult un element din S t = C . Adică funcția C : da injectiv, astfel încât dacă este aplicat la (s) T , atunci s S Este unic. Pentru ca codul să fie unic decodabil, C fiecare cuvânt sursă generează un alt cuvânt de cod decât celelalte cuvinte generate.

Injectivitatea funcției C : S 7 T nu asigură injectivitatea funcției C : da 7 T înțeles ca o concatenare a cuvintelor cod generate de simbolurile sursă fără goluri între ele. De exemplu, codul C folosit în codificarea rezultatelor meciurilor de fotbal nu poate fi decodat în mod unic, deoarece cuvântul de cod 00 poate fi decodat fie ca două victorii acasă, fie ca o victorie în deplasare.

00 = C „(1)C „(1) sau 00 = C „(2).

" " ′′ w3 = 0100011100011000100

Este evident că, dacă un simbol de separare între simboluri codificate este încorporat în șirul de cod, toate șirurile ar fi unic decodabile, totuși existența unor coduri unic decodabile fără a utiliza simboluri de separare este dovada că nu este necesar să se folosească un cuvânt de cod pentru a specifica distanța dintre fiecare simbol, cu o reducere corespunzătoare a dimensiunii cuvântului de cod final.

Dacă un cod este unic decodabil, orice cuvânt C poate fi descompus în mod unic.
unic ca o succesiune de cuvinte cod. Această caracteristică stă la baza următoarei teoreme/definiții.

Teorema 2.1 Dacă ui și vj reprezintă cuvinte de cod, ui = C(s) și vj = C(t) cu s,t S, următoarele două condiții sunt echivalente:

a) Codul C este unic decodabil.

T b) u1u2 ···un, v1v2 ··· vm C (S) u1u2 ···un = v1v2 ··· vm n = my i = 1,2,...,n vi = ui .

O modalitate de a folosi implicit simbolurile separatoare este prin a solicita ca simbolurile sursă să fie codificate prin cuvinte de cod de aceeași dimensiune. Astfel, dacă lungimea fiecărui cuvânt de cod al unui simbol sursă este n, toate cuvintele de cod vor avea o lungime care este un multiplu de n și decodificarea va fi efectuată în mod unic prin tăierea șirului primit în blocuri de n elemente și folosind injectivitatea lui. funcția C: S 7 T
. Un cod bloc este unul ale cărui cuvinte cod au toate aceeași lungime. Toate codurile de bloc sunt decodabile în mod unic.

Teorema 2.2 Dacă toate cuvintele de cod ale lui C : S 7 au aceeași lungime, deci C este decodificare
T blează în mod unic.

Reversul acestei teoreme nu este adevărat, deoarece există coduri unic decodabile ale căror cuvinte nu au aceeași lungime. Cuvintele codului C pentru piscine nu au aceeași lungime și sunt decodabile în mod unic, deși lungimea simbolului 1 îndeplinește o activitate similară ca separator:

- Dacă șirul începe cu 0, primul rezultat este o victorie a echipei gazdă și toate zerourile următoare sunt decodificate ca o victorie a echipei gazdă până când se ajunge la primul 1.
- Dacă șirul începe cu 1 și urmează un 0, perechea este decodificată ca o cravată.
- Dacă lanțul începe cu 1 și urmează un 1, perechea este decodificată ca o victorie în deplasare.
- Există șiruri care nu pot fi primite, cum ar fi 01.

Pentru a determina operațional dacă un cod este decodabil într-un mod unic este necesară existența unei proceduri operaționale care să ofere informații specifice în acest sens. Următoarele seturi servesc acestui scop.

Din codul C și recursiv, următorul lanț de mulțimi este definit:

- Dacă n = 0, atunci C0 = C.
- Dacă n N și n > 0, atunci

$$C_n = \{w \quad T^+ : uw = v \text{ unde } u \quad C \text{ și } v \quad C_{n-1} \text{ sau } u \quad C_{n-1} \text{ și } v \quad C\}.$$

- Dacă n = , C $= \bigcup_{n=1} C_n$.

1 După cum sa menționat anterior, este un abuz de notație, deoarece cu C ne referim la imaginea funcției injective C : S 7 T
.

Intuitiv, elementele lui Cn sunt determinate:

1. Calcularea cuvintelor lui Cn 1 care au ca prefix un cuvânt din codul C și cuvintele lui Cod C prefixat cu un cuvânt din mulțimea Cn 1.

2. Sufixele cuvintelor calculate mai sus sunt elementele lui Cn.

Deoarece C0 = C, cuvintele multimii C1 sunt sufixele cuvintelor codului C care au ca prefix un cuvânt C:

$$C1 = \{w \quad T^{+} : uw = v \text{ unde } u, v \quad C\}.$$

Pentru codul C ''′ a piscinelor, C ''′ 0 = {0,10,11}, deoarece nu există un cuvânt C. ''′ admite ca prefix a un cuvânt din codul C însuși ''′ De asemenea, ''′ C = /0, i = 2,3,... C ''′ i = /0 și C ''′ = /0.

Având în vedere acest exemplu și analizând definiția mulțimilor Ci , Este clar că dacă un set este goale, următoarele sunt de asemenea goale.

Proprietatea 2.1 Fie C un cod. Dacă Cn = /0, atunci Cn+1 = /0.

Deoarece definiția fiecărui Cn depinde numai de C și Cn 1, dacă două seturi se potrivesc, succesorii lor se potrivesc și.

Proprietatea 2.2 Fie C un cod. k N (Ci = Cj) = (Ci+k = Cj+k) .

Definiția recursivă poate sugera un număr infinit de mulțimi diferite Ci , dar acest lucru nu este adevărat, secvența (Ci)i N este constantă de la un anumit termen sau periodic, așa cum se va verifica mai jos. De fapt, cuvinte de lungime strict mai mare decât cele considerate nu apar niciodată în fiecare iterație.

Proprietatea 2.3 Fie C un cod astfel încât lungimea cuvintelor sale să fie ℓ1, ℓ2,..., ℓq. Dacă w Cn, atunci

$$|w| \quad \max{}' \{\ell_1, \ell_2, \ldots, \ell_q\}.$$

Demonstrație:

Prin inductie pe n.

- Dacă n = 0, rezultatul este imediat deoarece C0 = C.
- Ipoteza inducției: se presupune că dacă w Cn0 , apoi |w| max´ {ℓ1, ℓ2,..., ℓq}.
- Dacă w Cn0+1, atunci are loc una dintre următoarele:
  - u C uw
    Cn0 , de lungimea unui cuvânt |w| |uw| max´ {ℓ1, ℓ2,..., ℓq} și –după definiția Dacă uw Cn0 . apoi –prin ipoteza inducției– |uw| max´ {ℓ1, ℓ2,..., ℓq}
  - u Cn0 uw C.
    Dacă uw C, atunci |uw| max´ {ℓ1, ℓ2,..., ℓq} și –după definiția lungimii cuvântului– |w| |ww| max´ {ℓ1, ℓ2,..., ℓq}.

În orice caz, inegalitatea este valabilă.

Dacă lungimea tuturor cuvintelor din orice mulțime Ci este mai mică decât o legătură independentă pe i, atunci toate Ci au un număr finit de cuvinte.

Corolarul 2.1 Fie C un cod. i = 0,1,2,... |Ci | < .

În special, dacă ℓ = max´ {ℓ1, ℓ2,..., ℓq}, iar codul C este r-ary, suma numărului de cuvinte din lungimi mai mici sau egale cu ℓ,

$$N = r + r \quad \ell\, 2 + +r \quad = \frac{r(r\ \ell \quad 1)}{r \quad 1},$$

este o dimensiune a cardinalului de Ci

$$i = 0,1,2,\ldots\ |C_i| \quad r \quad \frac{r(r\ \ell \quad 1)}{1}.$$

Corolarul 2.2 Fie C un cod. i = 0,1,2,... |Ci | N = $\frac{r(r\ \ell \quad 1)}{r \quad 1}$.

Deoarece există 2N submulțimi distincte cu N sau mai puține elemente, pot exista doar 2N mulțimi distincte Ci , atunci intre multimile C0,C1,...,C2 N trebuie sa existe cel putin doua coincidente.

Corolarul 2.3 Fie C un cod. i, j 0,1,...,2 N, i < j Ci = Cj .

Conform corolarului 2.3 și proprietății 2.2, șirul C0,C1,...,Cn,... fie se stabilizează dintr-o anumită poziție, fie se repetă periodic mulțimile, în așa fel încât dacă i este o poziție i obținut conform la Corolarul 2.3, C = Dacă k=1Ck.

Corolarul 2.4 Fie C un cod. i 0,1,...,2 N, C $= \bigcup_{k=1}^{i} C_k$.

În general, calculul mulțimilor Ci nu este la fel de simplu ca în cazul codului C "´ a piscinelor, dar nici de obicei nu este o treabă imposibilă. Dacă C este un cod binar, r = 2 și lungimea maximă a cuvintelor sale este 3, fiecare mulțime Cn va avea cel mult N = 2(2 3 1) = 14 cuvinte, iar în șirul C1,C2,..., seturile vor începe să se repete înainte de poziția p = 2. Seturile Ci 14 = 16384.

oferă o metodă operațională pentru a determina dacă un cod este decodabil în mod unic.

Teorema 2.3 Sardine-Paterson. Un cod C este unic decodabil dacă și numai dacă C C = /0.

Demonstrație (parțială):

Dovada Teoremei Sardines-Paterson este relativ greoaie din punct de vedere tehnic deoarece analizează diferite cazuri. Mai jos este o schiță în care apar ideile și raționamentele principale care susțin demonstrația detaliată.

'= ' Se argumentează prin reductio ad absurdum presupunând că C este unic decodabil și C C / = /0, existând un cuvânt cod, w, care aparține lui C și (să spunem) C2, astfel încât să existe există u C și există v C1 astfel încât uw = vo vw = u.

Dacă v C1, atunci prin definiție există r,s C astfel încât rv = s.

- Dacă uw = v cu u C și v C1, atunci șirul ruw = rv = s poate fi decodificat în cel puțin două moduri diferite ca un singur cuvânt de cod, s C, sau ca trei cuvinte de cod, r,u, v C .
- Dacă vw = u cu u C și v C1, atunci șirul ru = rvw = sw poate fi decodificat în cel puțin două moduri diferite ca cuvintele de cod r,u C sau cuvintele de cod, s,w C, nota dată că r/ = s și că u/ = w deoarece v nu este cuvântul nul.

În orice caz, presupunerea că există un cuvânt comun pentru C și C contrazice faptul că C este unic decodabil.

' =' Prin reductio ad absurdum, se presupune că C C = /0 și că există un șir de cod t T care poate decodificat în două moduri diferite, t = uv = rs cu u, v,r,s C, u/ = r și v/ = s.

Nu poate fi că |u| = |r| pentru că atunci u = r. Presupunem fără pierdere de generalitate că |u| > |r|, atunci există w T +, astfel încât u = rw, rezultând w C1 și s C2 deoarece s = wv și v C, deci s C C2 C C , care contrazice faptul că C și C sunt disjunctive.

În definiția codului decodabil unic, decodificarea unică este necesară doar pentru șiruri finite de cuvinte de cod. Decodificarea unică poate fi definită și prin adăugarea condiției mai puternice ca toate șirurile – finite sau infinite – formate din cuvinte de cod să fie decodabile în mod unic.

Teorema Even-Levenshtein-Riley arată că un cod finit sau infinit este decodabil într-un fel unic dacă și numai dacă C C = /0 și există un indice n astfel încât Cn = /0.

Exercițiul 2.1 Studiați dacă următoarele coduri binare

a) C = {110,001,011,101,1111,1100}.

b) C = {110,001,011,101,1110,1100}.

c) C = {110,001,011,100,1111,1100}.

sunt decodabile în mod unic.

Solu ie:

Teorema Sardines-Paterson ne permite să determinăm dacă un cod este unic decodabil din mulțimile Cn = {w T + : uw = v unde u C și v Cn 1 sau u Cn 1 și v C} .

a) Dacă C = {110,001,011,101,1111,1100}, atunci

▪

▪

▪

Deoarece singurul cuvânt din C1 are lungimea 1 și toate cuvintele din C au lungime mai mare de 1,

: uw = v unde u C1 și v C} = {01,11}.

▪

Deoarece cuvintele lui C2 au lungimea 2 și toate cuvintele lui C au lungime mai mare de 2,

: uw = v unde u C2 și v C} = {1,0,11,00}.

▪

Deoarece cuvintele lui C3 au lungime mai mică sau egală cu 2 și toate cuvintele lui C au lungime mai mare de 2,

: uw = v unde u C3 și v C} = {10,01,111,100,0,11,00,1}.

▪

Deoarece cuvintele lui C4 au lungime mai mică sau egală cu 3 și toate cuvintele lui C au lungime mai mare sau egală cu 3,

C4 = {w ... C} = {10,01,111,100,0,11,00,1} = C4.

- Deoarece C4 = C5, atunci n 4 Cn = {0,1,00,01,10,11,100,111} și

$$C = \bigcup_{n=1} Cn = \{0,1,00,01,10,11,100,111\}.$$

Deoarece C C = /0, codul C este unic decodabil.

b) Dacă C = {110,001,011,101,1110,1100}, atunci

- C0 = C,
- C1 = {w T + : uw = v unde u, v C} = {0}.
- C2 = {w T + : uw = v unde u C și v C1 sau u C1 și v C}.
  Deoarece singurul cuvânt din C1 are lungimea 1 și toate cuvintele din C au lungime mai mare de 1,

$$C2 = \{w \quad T^+ : uw = v \text{ unde } u \quad C1 \text{ și } v \quad C\} = \{01,11\}.$$

- C3 = {w T + : uw = v unde u C și v C2 sau u C2 și v C}.
  Deoarece cuvintele lui C2 au lungimea 2 și toate cuvintele lui C au lungime mai mare de 2,

$$C3 = \{w \quad T^+ : uw = v \text{ unde } u \quad C2 \text{ și } v \quad C\} = \{1,0,10,00\}.$$

- C4 = {w T + : uw = v unde u C și v C3 sau u C3 și v C}.
  Deoarece cuvintele lui C3 au lungime mai mică sau egală cu 2 și toate cuvintele lui C au lungime mai mare de 2,

$$C4 = \{w \quad T^+ : uw = v \text{ unde } u \quad C3 \text{ și } v \quad C\} = \{10,01,110,100,11,1\}.$$

Cuvântul 110 C4 S n=1Cn = C și, de asemenea, aparține codului C, deci

110 C C și C C / = /0,

concluzionarea că codul C nu este unic decodabil.

Dacă C = {110,001,011,101,1110,1100}, șirul 1100011110 poate fi descompus ca o concatenare a cuvintelor C în două moduri diferite:

1100 |011 |110 și 110 |001 |1110,

deci codul C nu este unic decodabil.

c) Dacă C = {110,001,011,100,1111,1100}, șirul 1100011100 poate fi descompus ca concatenare a cuvintelor C în două moduri diferite:

1100 |011 |100 și 110 |001 |1100,

deci codul C nu este unic decodabil.

Exercițiul 2.2 În codul ternar, C, cuvântul 012120120 poate fi descompus ca o concatenare a cuvintelor cod în două moduri diferite: 012120 |120 și 01 |212 |01 |20.

Să se determine un cuvânt w astfel încât w C C .

Solu ie:

Codul C este format, cel puțin, din următoarele cuvinte

C=012120,120,01,212,20,....

O parte din seturile asociate sunt

- $C_0 = C$,
- $C_1 = \{w \quad T^+ : uw = v$ unde $u, v \quad C\} = \{2120, \ldots\}$.
- $C_2 = \{w \quad T^+ : uw = v$ unde $u \quad C$ și $v \quad C_1$ sau $u \quad C_1$ și $v \quad C\}$. 2120 nu este

  un prefix al niciunui element din C_1, ci 212 $\quad C$ și 212|0 $\quad C_1$, atunci

$$C_2 = \{w \quad T^+ : uw = v \text{ unde } u \quad C_1 \text{ și } v \quad C\} = \{0, \ldots\}.$$

- $C_3 = \{w \quad T^+ : uw = v$ unde $u \quad C$ și $v \quad C_2$ sau $u \quad C_2$ și $v \quad C\}$.

  Deoarece singurul cuvânt localizat din C_2 are lungimea 1, nu există cuvinte C care să prefixeze acel cuvânt din C_2, deci

$$C_3 = \{w \quad T^+ : uw = v \text{ unde } u \quad C_2 \text{ și } v \quad C\} = \{12120, 1, \ldots\}.$$

- $C_4 = \{w \quad T^+ : uw = v$ unde $u \quad C$ și $v \quad C_3$ sau $u \quad C_3$ și $v \quad C\}$.

  Nu există cuvinte C care sunt un prefix al oricărui cuvânt C_3 localizat, dar există un cuvânt C_3 care este un prefix al oricărui cuvânt C localizat,

$$C_4 = \{w \quad T^+ : uw = v \text{ unde } u \quad C_3 \text{ și } v \quad C\} = \{20, \ldots\}.$$

Cuvântul $w = 20 \quad C \quad C$.

Este interesant de observat că dacă codul considerat ar avea doar cuvintele localizate, adică C = 012120,120,01,212,20, nu ar fi decodabil unic, ci C_5 = {0, ceea ce arată că nu este suficient ca șirul de seturi C_i este setată la setul gol, astfel încât codul să poată fi decodat în mod unic.

Exercițiul 2.3 Fie C un cod.

a) Analizați diferitele situații în care un cuvânt $w \quad C \quad C_3$.

b) Aplicați rezultatele obținute pentru a obține un șir cu două decodificări diferite ale codului C = {000,11,110,011}.

Solu ie:

a) Fie $W \quad C$.

Dacă $u \quad C_1$, atunci w', $w'' \quad C$ $w'u = w''$.

w aparține lui C_3 dacă oricare dintre următoarele este adevărată:

- $w_1 \quad C$, $v_1 \quad C_2$ $w_1 w = v_1$.

  Dacă $v_1 \quad C_2$ se datorează faptului că apare una dintre următoarele situații:

  - $w_3 \quad C$, $u_1 \quad C_1$ $w_3 v_1 = u_1$.

    Dacă $u_1 \quad C_1$, atunci $w_5, w_6 \quad C$ $w_5 u_1 = w_6$.

    Înlocuind cuvintele care nu aparțin lui C, în această ultimă egalitate, obținem $w_5 w_3 v_1 = w_6$

    Y

$$w_5 w_3 w_1 w = w_6.$$

    Cuvântul de cod w_6 poate fi descompus ca o concatenare a cuvintelor de cod w_5, w_3, w_1 și w.

- w4 C, u2 C1 u2v1 = w4.
  Dacă u2 C1, atunci w7,w8 C w7u2 = w8 deci w7u2v1 = w7w4 și w8v1 = w7w4. Înlocuind trecerea v1 cu concatenarea a două cuvinte, obțineți

  w8w1w = w7w4

  apoi concatenarea cuvintelor de cod w8, w1 si w produce sirul format din concatenarea cuvintelor w7 si w4.
  w2 C, v2 C2 v2w = w2.

■

Dacă v2 C2, se datorează faptului că apare una din următoarele situații:
- w9 C,u3 C1 w9v2 = u3.
  Dacă u3 C1, atunci w11,w12 C w11u3 = w12.
  Concatenând prima egalitate cu w9 , obținem w9v2w = w9w2 și u3w = w9w2 și concatenând această ultimă egalitate cu w11, obținem w11u3w = w11w9w2 sau altfel

  w12w = w11w9w2

  atunci concatenarea cuvintelor w12 și w este egală cu concatenarea cuvintelor w11, w9 și w2.
- w10 C,u4 C1 u4v2 = w10.
  Dacă u4 C1, atunci w13,w14 C w13u4 = w14, deci w13u4v2 = w13w10 și w14v2 = w13w10.
  Concatenând w14 în egalitatea v2w = w2, obținem w14v2w = w14w2 și

  w13w10w = w14w2

  atunci concatenarea cuvintelor w13, w10 și w este egală cu concatenarea cuvintelor w14 și w2.

b) Dacă C = {000,11,1100,011}, atunci C0 = C, C1 = {00}, C2 = {0}, C3 = {00,11}, C4 = {0,00}, C5 = {0,00,11}, C6 = {0,00,11}.

Al 11-lea cuvânt C C3.

Cunoscând C2, se poate asigura că 0 |11 = 011 C cu 0 C2.

Având în vedere C1, 00 |0 = 000 C cu 00 C1 care provine din 11 |00 = 1100 C cu 11 C.

Concatenând 0 |11 = 011 cu 00, obținem 00 |0 |11 = 00 |011 care dă un cuvânt de cod din 0 C2 grupând primele două elemente, 000 |11 = 00 |011. Prin concatenarea acestui șir rezultat cu cuvântul 11, se formează un cuvânt din 00 C1, 11 | 000 |11 = 11 |00 |011 în așa fel încât să existe un șir de cuvinte de cod care pot fi decodificate în două moduri:

11 |000 |11 = 1100 |011.

Exercițiul 2.4 Verificați dacă codul ternar C = {02,12,120,21} este decodabil unic, dar există șiruri infinite care pot fi decodabile în două moduri diferite.

Solu ie:

Seturile asociate codului C sunt

C0 = C, C1 = {0}, C2 = {2}, C3 = {1}, C4 = {2,20}, C5 = {1}, C6 = {2,20},... și C = {0, 1,2,20}

deci C C = /0 și, după teorema Sardines-Paterson, C este unic decodabil.

Deoarece nu există un index din care mulțimile asociate codurilor să fie goale, ipotezele teoremei Even-Levenshtein-Riley nu sunt îndeplinite și există cel puțin un lanț infinit care poate fi decodificat în două moduri diferite:

120212121... = 120 |21 |21 |21 | ... = 12 |02 |12 |12 |12....

3. Coduri instantanee și presetări

Exemplul 3.1 Codul binar C = {0,01,011,111} este decodabil unic deoarece

$$C_0 = C,\ C_1 = \{1,11\} \text{ și } C_2 = \{11,1\}.$$

În general, pentru a putea decoda un mesaj codificat cu C, trebuie să așteptați până când primiți secvența completă de biți, deoarece o secvență care începe de la 0 și apoi are uni, $0111\cdots11$, depinde de numărul de biți. cele pe care trebuie să le descompună Într-un fel sau altul:

- Dacă sunt 0, descompunerea este banală: 0.
- Dacă există 1 unul, descompunerea este 01.
- Dacă sunt 2, descompunerea este 011.
- Dacă sunt 3, descompunerea este $0|111$.
- Dacă sunt 4, descompunerea este $01|111$.
- Dacă sunt 5, descompunerea este $011|111$.
- Dacă sunt 6, descompunerea este $0|111|111$.

Deci, dacă există k,

- Dacă k = 3n, atunci $01\cdots1 = 0|\underbrace{111|\cdots|111}_{\frac{k}{3}}$.
- Dacă k = 3n+1, atunci $01\cdots1 = 01|\underbrace{111|\cdots|111}_{\frac{k-1}{3}}$.
- Dacă k = 3n+2, atunci $01\cdots1 = 011|\underbrace{111|\cdots|111}_{\frac{k-2}{3}}$.

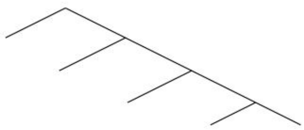
Dacă șirul primit începe cu 1, primele 3 trebuie să fie codificate ca 111, indiferent a informațiilor ulterioare.

Suntem interesați să găsim coduri unic decodabile a căror decodare se poate face instantaneu. fără a fi nevoie să aștepte să primească toate simbolurile mesajului.

Un cod C este instantaneu, dacă pentru orice secvență de cuvinte de cod, $w = w_1w_2 \ldots$, orice subsecvență care începe la fel cu w este decodificată în mod unic, indiferent de cuvintele ulterioare.

Codul C din exemplul 3.1 nu este instantaneu. Codul C are$'$ = {0,10,11} este instantaneu deoarece există numai trei moduri de decodare și nu depinde de cuvintele de la sfârșitul șirului:

- Dacă apare un 0, cuvântul primit poate fi decodat doar ca 0.
- Dacă apare un 1, cuvântul trebuie să aibă doi biți:
  - Dacă este 0, cuvântul primit este 10.
  - Dacă este 1, cuvântul primit este 11.



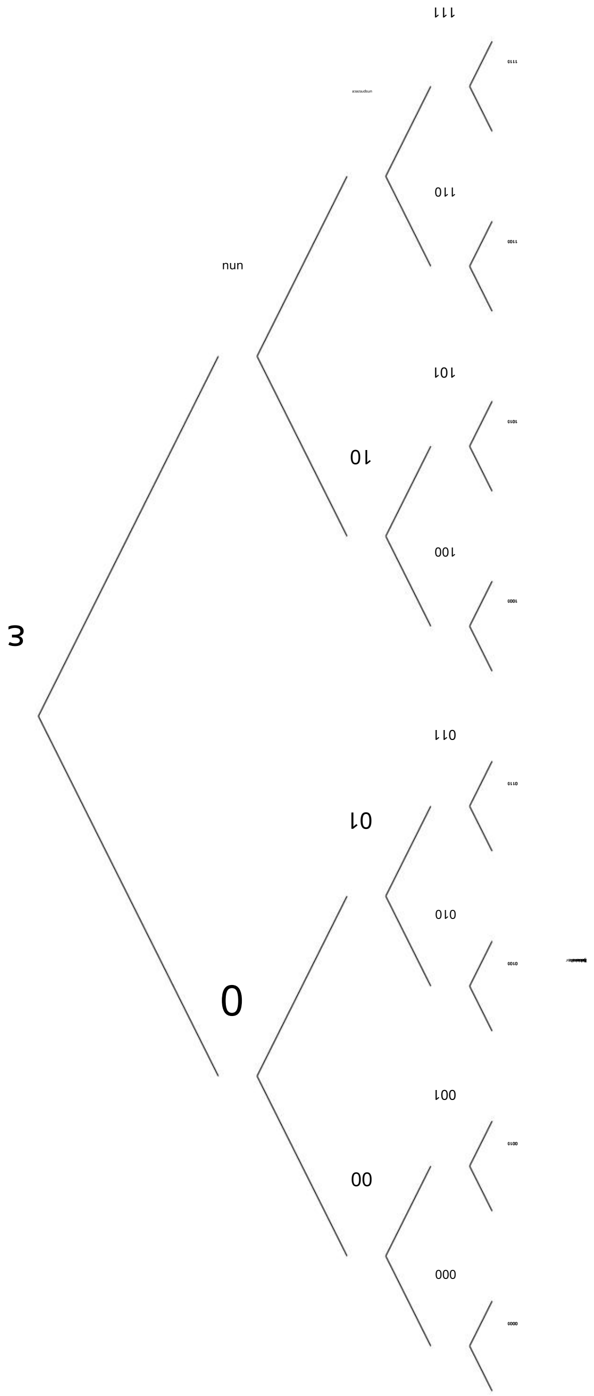
$\bar{r}$

$\overline{1}\,2$

$\overline{1}\,4$ $\overline{1}\,4$

$$\sum_{i=1}^{p} \frac{}{\ell_i} \quad 1.$$

[2] $n\,r$ reprezintă conceptul imprecis de proporție a unui arbore sub un vârf etichetat cu un cuvânt de lungime n. Nu sunt uitați că a priori adâncimea arborelui spanning nu este limitată.


0
00
01
10
000
001
010
011
100
101
110
111

Există coduri instant ternare cu cinci cuvinte de lungimi 1,2,3,3,4 de atunci

$$\frac{1}{3}+\frac{1}{3^2}+\frac{1}{3^3}+\frac{1}{3^3}+\frac{1}{3^4}=\frac{43}{81}\leq 1,$$

de exemplu, C = {0,10,110,120,2222}, dar nu există coduri instantanee binare cu cinci cuvinte cu lungimea 1,2,3,3,4, deoarece

$$\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\frac{1}{2^3}+\frac{1}{2^4}=\frac{17}{16}>1.$$

4. Inegalități Kraft și McMillan

Relația dintre lungimile cuvintelor unui cod și faptul că este instantaneu este mai puternică decât ceea ce a fost analizat până acum. Inegalitatea lui Kraft oferă o condiție necesară și suficientă pentru existența codurilor instantanee în funcție de lungimea cuvintelor lor.

Teorema 4.1 Există un cod r-ary instantaneu cu lungimi de cuvinte l1,...,lq dacă și numai dacă

$$\sum_{i=1}^{q}\frac{1}{r^{l_i}}\leq 1.$$

Demonstrație:

'= ' Fie C un cod instant sau prefix și $\ell = \max\{l_1,\dots,l_q\}$.

În arborele complet, nodurile de la nivelul ℓ pot să nu provină din niciun cuvânt de cod sau pot proveni doar dintr-un cuvânt de cod cel mult, deoarece dacă ar exista două cuvinte de cod diferite care conduc la un cuvânt de nivel superior, atunci unul ar fi prefixul celuilalt ℓ ℓ_i nodurile de nivel ℓ, care corespund Din fiecare cuvânt de cod w_i de lungime ℓ_i r poate fi accesat cu toate cuvintele care încep cu w_i și având lungimea ℓ.

În total, din cuvintele de cod, puteți accesa $\sum_{i=1}^{q} r^{\ell-\ell_i}$ cuvinte de la nivelul ℓ. Ca și în nivelul ℓ exista r^{ℓ} cuvinte, atunci

$$\sum_{i=1}^{q} r^{\ell-\ell_i}\leq r^{\ell}.$$

Împărțirea la r^{ℓ}, obținem inegalitatea dorită

$$\sum_{i=1}^{q} r^{-\ell_i}=\sum_{i=1}^{q}\frac{1}{r^{\ell_i}}\leq 1.$$

' =' Fie C un cod cu q cuvinte de lungime $\ell_1, \ell_2,\dots, \ell_q$ astfel încât $\sum_{i=1}^{q}\frac{1}{r^{l_i}}\leq 1.$

Fără pierderea generalității se poate presupune că $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_q$.

Deoarece nu există cuvinte cu lungime mai mare de ℓ_q, luăm în considerare arborele de acoperire finit al ℓ_q cuvintelor cu cuvântul gol și se termină la nivelul ℓ_q. În total există r cod de lungime maximă: toate cuvintele de lungime ℓ_q care de cod candidate începând pot fi formate cu r simboluri.

Luăm orice cuvânt de lungime ℓ_1, w_1 și tăiem arborele în așa fel încât toate ramurile care duc la cuvintele de nivel ℓ_q care încep cu w_1 să fie eliminate ℓ_q ℓ_1. S-au eliminat cuvintele r de nivelul ℓ_q, toate de lungime ℓ_q și care încep cu w_1, împreună cu cele de lungime între ℓ_1 și ℓ_q începând cu w_1.

Dacă q = 1, codul rezultat este instantaneu, deci rezultatul este imediat.

Dacă q > 1, folosind ipoteza teoremei, obținem inegalitatea

$$r^{\ell_q - \ell_1} = r^{\ell_q}\frac{1}{r^{\ell_1}} < r^{\ell_q}\sum_{i=1}^{q}\frac{1}{r^{l_i}} \le r^{\ell_q}$$

ceea ce asigură că la nivelul ℓ_q mai există cuvinte care nu au ca prefix w_1 .

În ramurile care nu au fost tăiate trebuie să existe un cuvânt, w_2, de lungime $\ell_2 \ge \ell_1$. Luăm w_2 ca al doilea cuvânt al codului și tăiem arborele eliminând toate ramurile care conduc la $r^{\ell_q - \ell_2}$ cuvinte mai puțin la nivelul ℓ_q al arborelui. înseamnă r În acești doi pași au fost eliminate cuvintele de nivelul ℓ_q care încep cu w_2, ceea ce

$$r^{\ell_q - \ell_1} + r^{\ell_q - \ell_2} = r^{\ell_q}(r^{-\ell_1} + r^{-\ell_2}) = r^{\ell_q}\sum_{i=1}^{2} r^{-\ell_i}$$

cuvinte cu lungimea ℓ_q deoarece, având ordonate lungimile, nu există ramuri comune pentru w_1 și w_2 și nici w_2 nu are forma $w_2 = w_1 w'$.

Dacă q = 2, procesul este încheiat și avem un cod prefix.

Dacă q > 2, procesul poate fi continuat deoarece, prin ipoteză,

$$r^{\ell_q}\sum_{i=1}^{2} r^{-\ell_i} < r^{\ell_q}\sum_{i=1}^{q} r^{-\ell_i} \le r^{\ell_q}$$

și mai există cuvinte la nivelul ℓ_q care nu încep cu w_1 sau w_2.

Repetând procesul q − 1 ori, ajungem la inegalitate

$$r^{\ell_q - \ell_1} + r^{\ell_q - \ell_2} + \dots + r^{\ell_q - \ell_{q-1}} = r^{\ell_q}\sum_{i=1}^{q-1} r^{-\ell_i} < r^{\ell_q}\sum_{i=1}^{q} r^{-\ell_i} \le r^{\ell_q}$$

ceea ce asigură că mai există la nivelul ℓ_q un cuvânt care nu începe cu niciunul dintre cuvintele q − 1 alese. Acest cuvânt împreună cu cele precedente formează un prefix și un cod instant.

Fiecare cod instant este unic decodabil, dar există coduri unic decodabile care nu sunt instantanee. S-ar părea rezonabil că o condiție necesară și suficientă pentru determinarea existenței codurilor unic decodabile pe baza lungimii cuvintelor lor ar fi mai relaxată decât cea oferită de inegalitatea lui Kraft pentru codurile instant. Totuși nu este așa. Inegalitatea lui McMillan stabilește aceeași condiție pentru ambele tipuri de coduri.

Teorema 4.2 Există un cod r-ary decodabil unic cu lungimi de cuvinte $l_1,\dots,l_q$ dacă și numai dacă

$$\sum_{i=1}^{q}\frac{1}{r^{l_i}} \le 1.$$

Dovada se găsește la paginile 15 și 16 din Informația și codificarea Tehory de Gareth A. Jones și J. Mary Jones.

Din inegalitățile Kraft și McMillan se deduce următorul rezultat:

Corolar 4.1 Există un cod r-ary decodabil unic cu lungimea cuvântului $l_1,\dots,l_q$ dacă și numai dacă există un cod r-ary instantaneu cu lungimea cuvântului $l_1,\dots,l_q$

Aceste inegalități afirmă că există coduri cu anumiți parametri care sunt instantanee și codabile într-un mod unic și trebuie luate în considerare următoarele precizări:

- Dacă un cod este instantaneu, este deja decodabil în mod unic, atunci, dacă construiți un cod instantaneu, aveți deja unul decodabil în mod unic, cu aceleași lungimi de cuvinte.
- Inegalitățile Kraft și McMillan afirmă că codurile instantanee și decodificabile pot fi găsite numai dacă sunt îndeplinite condițiile, dar nu spun că un cod care îndeplinește acele condiții este unic instantaneu sau decodabil. De exemplu, codul C = {0,00,000} nu este decodabil în mod unic, chiar dacă lungimile cuvintelor sale satisfac inegalitatea necesară.
- Codurile decodificabile unic nu trebuie să fie instantanee, dar cu siguranță există un cod instantaneu care îndeplinește condițiile de lungime a cuvântului. De exemplu, codul C1 = {0,01,11} nu este instantaneu și codul C2 = {0,10,11}, cu aceleași lungimi de cuvinte, este instantaneu.

În tabelul de mai jos

Coduri binare cu lungimi de cuvinte 1, 2 și 1, 2, 2 {0,00} {1,11}		
		{0,00,01} {0,00,10} {0,00,11} {0,01 . 10} {1,11,10} {1,11,01} {1,11,00} {1,10,01} {0,01,11} {1,10,00} {0,10, 11 } {1,01,00}
Decodabil numai din	Fără prefixe	{0,01} {1,10}
	prefixe	{0.10} {0.11} {1.01} {1.00}

cele 8 coduri binare ale cuvintelor lungimii 1, 2 și cele 12 coduri binare ale cuvintelor lungimii 1, 2, 2 sunt clasificate în funcție de faptul că sunt sau nu decodabile unic. La rândul lor, cele decodabile unic sunt clasificate în prefixe și non-prefixe. Se poate observa că atunci când este îndeplinită condiția inegalităților Kraft și McMillan, în ambele cazuri există coduri de prefix și coduri decodabile unic care nu sunt prefixe.

Exercițiul 4.1 Câte coduri ternare instantanee există cu 9 cuvinte de lungime 1,2,2,2,2,2,3,3,3?

Solu ie:

Privind arborele de generare a cuvintelor de cod din Tabelul 2, există 3 moduri de a alege un cuvânt de lungime 1.

6
5 Alegând un cuvânt cu lungimea 1, puteți alege până la 6 cuvinte cu lungimea 2, ceea ce face un total de opțiuni dacă doriți să alegeți 5 cuvinte.

Au fost alese 5 cuvinte cu lungimea 2, au mai rămas doar 3 cuvinte cu lungimea 3 pentru care să fie alese indeplinesc conditiile declarației. 6 În total,
sunt 3 5
= 18 coduri ternare instantanee cu 9 cuvinte de lungimi 1,2,2,2,2,2,3,3,3.

Exercițiul 4.2 Câte secvențe de cod de lungime ℓ, Nℓ , pot fi formate cu cuvintele de cod C = {0,10,11}?

Solu ie:

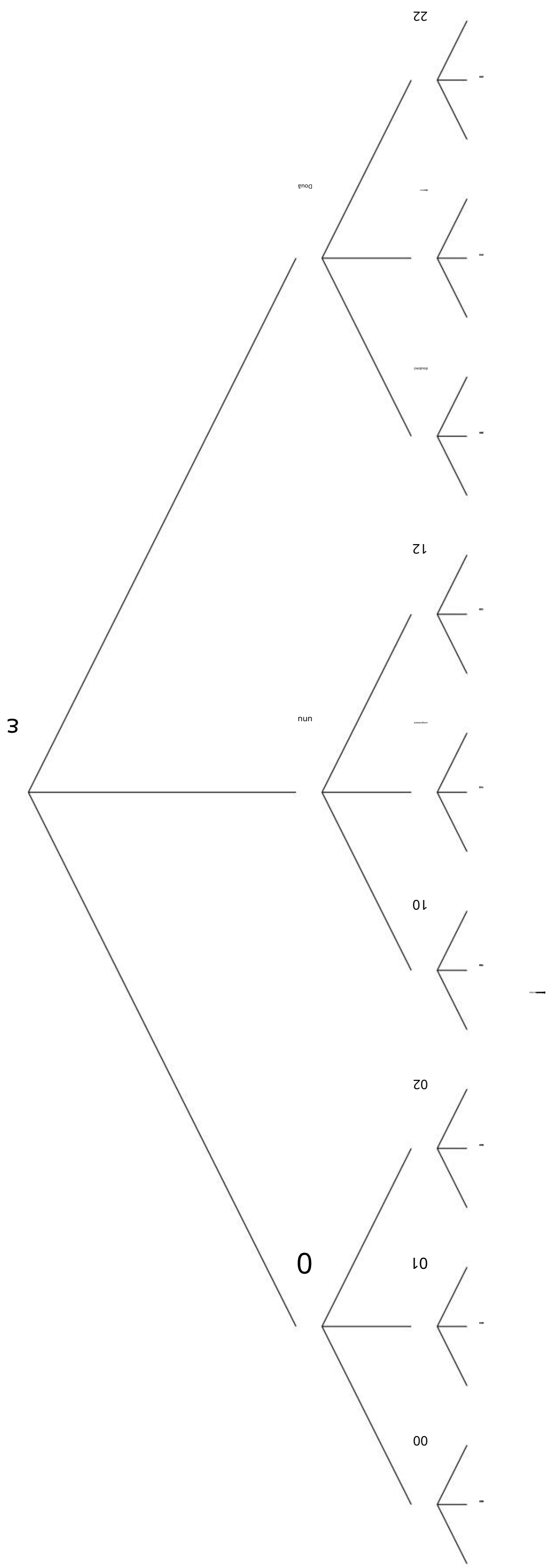
Dacă ℓ = 1, există o singură secvență de cod, deoarece există un singur cuvânt cu lungimea 1, deci N1 = 1.

Dacă ℓ = 2, există doar trei secvențe de cod: 00, 10 și 11, atunci N2 = 3.

Dacă ℓ = 3, deoarece lungimea cuvintelor de cod este mai mică sau egală cu doi, secvența trebuie să fie alcătuită din cel puțin două cuvinte, s = w, cu w C. Cu ultimul cuvânt al secvenței w = 0 , pot fi formate secven e N2 ; cu ultimul cuvânt se pot forma w = 10 N1 secvențe și cu ultimul cuvânt se pot forma w = 11 N1 cuvinte, astfel încât N3 = N2 +N1 +N1.

Dacă ℓ = 4, fie cuvântul 0 a fost adăugat la sfârșitul unei secvențe de lungime 3, fie un cuvânt de lungime 2 (10 sau 11) la sfârșitul unei secvențe de lungime 2, apoi N4 = N3 +2N2.

În general, Nℓ satisface relația recursivă Nℓ = Nℓ 1 +2Nℓ 2 cu N2 = 3 și N1 = 1.



$N_n = A2^n + B(-1)^n$

3 și B =

ℓ

ce

ℓi 1, se pot forma? r

i=1

sub al doilea cuvânt ales: r r r de alegerea celui de-al treilea cuvânt. ℓ3 ℓ3 ℓ1 ℓ3 ℓ2 cele de ℓ2 r

ℓ1+ +ℓq 1 r ℓ1 1 r ℓ1 ℓ2 r 1 r ℓ1 ℓ2 r ⋯ r ℓq 1

1 1 2 3

1+1 1

5. Coduri cuprinzătoare

Machine Translated by Google

Într-un snapcode de lungime maximă ℓ, nu toate șirurile de lungime ℓ sunt prefixate de un cuvânt de cod. Într-un cod snap, pot exista secvențe de cuvinte de lungime maximă ℓ care sunt prefixate de unul sau mai multe cuvinte de cod.

Proprietatea 5.2 Dacă un cod r-ary cu cuvinte de lungimi ℓ1, ℓ2,..., ℓq este exhaustiv, atunci

$$\sum_{i=1}^{ce} \ell_i \quad 1.\ r$$

Demonstrație:

Dacă C este exhaustiv de nivelul ℓ, atunci numărul de secvențe de lungime ℓ care sunt prefixate cu un cuvânt cod, $\sum_{i=1}^{ce} r^{\ell \ \ell_i}$, trebuie să fie mai mare decât numărul de cuvinte de lungime ℓ, r^{ℓ}, astfel încât

$$\sum_{i=1}^{ce} r^{\ell \ \ell_i \ \ell} \ r \ .$$

Împărțirea la r^{ℓ}, se obține inegalitatea dorită, $\sum_{i=1}^{ce} r^{\ell_i}$ 1.

da $\sum_{i=1}^{ce}$ ℓi = 1, atunci fiecare cuvânt de lungime ℓ este prefixat cu unul și numai un cuvânt de cod, r

deci C este prefix și instantaneu. În schimb, dacă C este instantaneu și $\sum_{i=1}^{ce}$ ℓi = 1, deci C este evacuarea r deoarece fiecare cuvânt de lungime ℓ nu poate proveni din mai mult de un cuvânt de cod, suma tuturor cuvintelor care provin din cuvintele de cod este exact r^{ℓ}.

Corolarul 5.1 Fie C un cod r-ary cu cuvinte de lungimi ℓ1, ℓ2,..., ℓq. Este împlinită

a) Dacă C este exhaustiv i $\sum_{i=1}^{ce}$ ℓi = 1, deci este instantaneu. r

b) Dacă C este instantaneu i $\sum_{i=1}^{ce}$ ℓi = 1, deci este exhaustiv. r

c) Dacă C este instantaneu și exhaustiv, atunci $\sum_{i=1}^{ce}$ ℓi = 1. r

Cu toate acestea, relațiile de mai sus nu sunt echivalente din moment ce

- Există coduri instant, cum ar fi C = {0} care nu este exhaustiv și $\sum_{i=1}^{ce}$ ℓi = 2r 1 / = 1.
- Există coduri exhaustive, cum ar fi C = {0,1,00} care nu este instantanee și $\sum_{i=1}^{ce} r^{\ell_i}$ = 5/ = 1. 4
- Există coduri, cum ar fi C = {0,00,10} care se întâlnesc $\sum_{i=1}^{ce}$ ℓi 1 2 2 = 2 +2 +2r = 1 și care nu sunt nici evacuare tive sau instantanee.