COMPRESIA DATELOR - GENERALITĂȚI

1. Definiții

Termenul de compresie a datelor este, sper, unul destul de cunoscut. O posibilă definiție generală ar fi:

> ***Compresia de date** este ansamblul prelucrărilor ce se aplică unor date în scopul reducerii dimensiunii reprezentării acestora*.

Eficiența compresiei obținută cu o metodă oarecare poate fi apreciată prin raportul de compresie. În contextul definiției anterioare

> ***Raportul de compresie** este egal cu raportul dintre dimensiunea reprezentării datelor în lipsa compresiei şi dimensiunea reprezentării datelor obținute în urma compresiei.*

Uneori, pentru rapoarte de compresie mici, se consideră raportul invers, exprimat de cele mai multe ori în procente. De obicei trebuie determinat, în funcție de context, care din cele două abordări este cea folosită. În diverse cazuri particulare raportul de compresie este redefinit (particularizat) corespunzător situațiilor respective.

Compresia se realizează prin schimbarea modului de reprezentare a datelor având deci de a face cu un caz particular de **codare** (de aceea termenii de compresie şi de codare vor fi folosiți aproape ca sinonime, dupa cum sunt consacrați pentru diferite metode). Codarea se face în raport cu un anumit model al datelor, aflându-ne astfel în cazul mai general de **modelare**, al căutării unui model corespunzător (optim) al datelor.

2. Clasificări

În cadrul acestui domeniu există o mare varietate de metode, de cele mai multe ori prezentate individual, fără a se încerca integrarea acestora. Acest lucru impune clasificări, care se pot face după diverse criterii.

Cea mai generală clasificare a metodelor de compresie se face după **eroarea de refacere** a datelor. În raport cu acest criteriu distingem două categorii mari de metode:

1. Metode fără pierderi - în care datele se refac în totalitate, fără a exista nici o diferență între datele originale şi cele refăcute. Aceste metode sunt cunoscute ca şi metode **"LOSSLESS"**. Ele se aplica în general în situațiile în care refacerea fără eroare este esențială: compresie de programe executabile, surse de programe, texte, în general date de natură strict numerică.

2. Metode cu pierderi - în care datele se refac în limita unor erori considerate acceptabile. Acceptarea din start a existenței unor erori duce la creşterea spectaculoasă a ratei de compresie față de cazul anterior. Termenul consacrat pentru aceste metode este cel de metode **"LOSSY"**. Aceste metode se aplică în general în cazul datelor de natură analogică (date ce reprezintă semnale care la origine erau analogice) destinate în final tot unui "operator" uman. În această categorie intră compresia semnalului audio şi a semnalului video, semnale în a căror prelucrare, datorită naturii lor analogice, găsim de obicei un cuantizor care se constituie oricum într-o sursă de erori.

Delimitarea unei metode generale în lossless / lossy nu este strictă. De exemplu, tehnicile predictive se pot încadra în ambele categorii, în funcție de datele pentru care se aplică (de natură numerică sau analogică) şi în funcție de existența în algoritm a unui cuantizor.

În funcție de **modul cum evolueaza în timp modelul sursei** de date, metodele de compresie se pot împărți astfel:

1. Metode statice - în care modelul este fix, nu evoluează în timp ci este construit apriori, pe baza unor mesaje considerate tipice. În această categorie este inclusă compresia realizată în cazul transmisiilor fax. Este utilă în situațiile în care cunoaştem statistica fluxului de date şi aceasta ramâne neschimbată, rezultând metode dedicate unui anumit context. Datorită caracterului static se pretează bine chiar şi la implementări hardware. Schemă generală de compresie (codare) pentru acestă situație este dată în Figura 1.1.

2. Metode semistatice - în care modelul este construit înaintea codării (comprimării) pe baza datelor

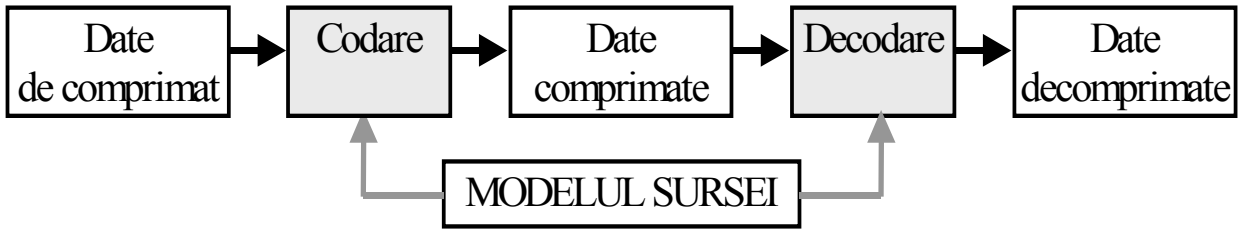

Figura 1.1 Schema generală de compresie pentru modelare statică

ce urmează să fie comprimate şi este menținut nemodificat pe durata acesteia. Avantajul constă în adaptarea metodei la fluxul de date (metodele fiind mai general aplicabile). Dezavantajele sunt necesitatea parcurgerii de două ori a fluxului de date şi necesitatea transmiterii modelului spre decodor, fapt care încarcă suplimentar fluxul de date şi deteriorează performanțele. În această categorie se încadrează metoda de comprimare bazată pe algoritmul Huffman (cunoscută de obicei ca şi metoda Huffman static – notație oarecum nefericită dar consacrată). Schemă generală de compresie (codare) pentru această situație este dată în Figura 1.2.

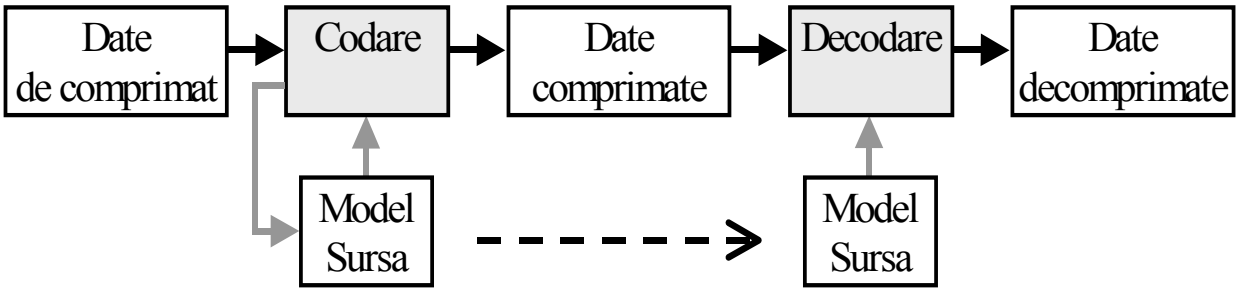

Figura 1.2 Schema generală de compresie pentru metodele semistatice

3. Metode dinamice (adaptive) - în care compresia începe la o anumită stare a modelului, aceeaşi atât în cazul comprimării cât şi în cel al decomprimării. Starea inițială a modelului poate fi la rândul ei determinată ca şi în unul din cazurile anterioare. Fiecare simbol se codează pe baza modelului curent. După codare sau respectiv decodare (la decodor) starea modelului se actualizează în acelaşi mod atât la codor cât şi la decodor. Se elimină astfel cele două dezavantaje importante ale metodelor semistatice. Cele mai cunoscute metode de acest tip sunt compresia Huffman adaptivă şi metodele de dicționar. În acest caz schema generală de compresie devine cea din Figura 1.3.

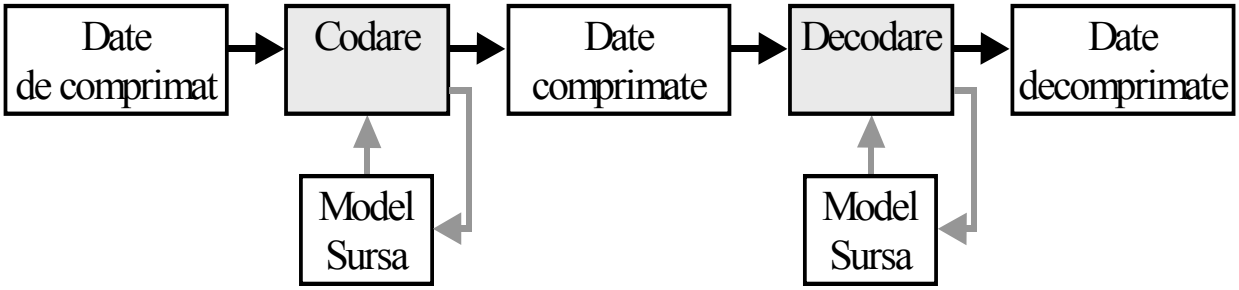

Figura 1.3 Schema generală de compresie pentru modelare dinamică (adaptivă)

În funcție de **natura informațiilor** pe baza cărora se construieşte modelul putem clasifica metodele de compresie în (clasificare care acoperă doar metodele lossless):

1. Metode de compresie bazate pe modelarea statistică a surselor discrete

Modelarea statistică a unei surse discrete de informație constă în asocierea unei probabilități de apariție fiecărui simbol al alfabetului sursei urmată apoi de atribuirea de cuvinte de cod cu un număr mai mic de biți cuvintelor cu o probabilitate de apariție mai mare. Din această categorie fac parte codificarea Shannon-Fano, codificarea Huffman statică (optimă pentru cazul codificării simbol cu simbol), codificarea Huffman dinamică, codificarea aritmetică (codifică secvențe de simboluri şi nu simboluri individuale).

2. Metode de compresie bazate pe modelarea lingvistică a surselor discrete

Aceste metode au la bază faptul că şirul de simboluri generate de sursă prezintă constrângeri în ceea ce priveşte combinațiile care pot să apară. Datorită acestor constrângeri nu toate combinațiile posibile apar cu aceeaşi frecvență. Ideea este de a construi un dicționar din cuvinte

ale limbajului asociat sursei. Ulterior inserării unui cuvânt în dicționar în cazul oricărei apariții a cuvântului acesta se înlocuieşte cu o informație despre poziția sa în dicționar. În acest caz modelul sursei este reprezentat de dicționar. Din această categorie fac parte algoritmii LZ77, LZ78 , LZW.

3. Model general de compresie - decompresie

Cel mai general model al unui sistem de compresie de date este cel din Figura 1.4. Prima transformare este transformarea care reduce entropia cu scopul de a elimina parametrii mesajului care nu au importanță pentru destinatar (specific cazului lossy). Transformarea următoare conservă entropia datelor dar reduce redundanța în scopul măririi eficienței transmisiunii sau stocării (specific cazului lossless). În cazul în care transmisia se face pe un canal cu zgomot se poate reintroduce o anumită redundanță prin utilizarea de coduri corectoare sau detectoare de erori. Deoarece aceste coduri nu realizează o compresie ci dimpotrivă o creştere a dimensiunii datelor ele nu vor fi tratate în continuare.

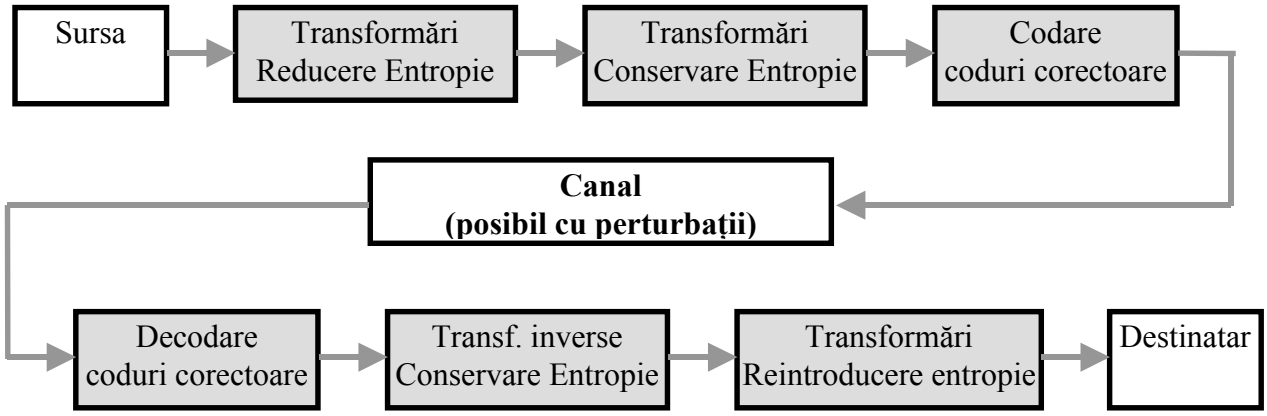

Figura 1.4 Modelul general de compresie - decompresie

4. Câteva exemple

În continuare prezentăm câteva exemple simple, relativ empirice, de metode de compresie. Acestea nu se folosesc în practică de sine stătător, în această formă simplă, ci de obicei ca etape intermediare în scheme mai elaborate. Scopul principal al prezentării acestora este acela de a exemplifica conceptele de metodă statică / semistatică / dinamică şi de metodă statistică / de dicționar. Analizând rezultatele obținute vom observa ca niciuna din metode nu este utilă în toate situațiile, pentru unele fluxuri (nefericit alese) nu doar că nu obținem compresie, ci am putea avea chiar o creştere a dimensiunii datelor.

4.1. Codare RLE (Run Length Encoding)

Prezentăm în continuare o versiune RLE aşa cum este implementată în formatul BinHex 4.0. Acest format impune, pe lângă codarea RLE, şi existența unui comentariu standard, unui antet impus şi

codare pe 7 biti a fluxului binar rezultat. În cele ce urmează prezentăm doar etapa RLE, celelalte pot fi găsite în documentația BinHex 4.0.

Ideea de bază la RLE este aceea de codare a lungimii secvențelor repetitive (numite "runs"). În continuare prezentăm un flux de intrare şi fluxul rezultat în urma codării RLE:

Flux intrare	Flux codat RLE
00 11 22 33 44 55 66 77	00 11 22 33 44 55 66 77
11 22 22 22 22 22 22 33	11 22 90 06 33
11 22 90 33 44	11 22 90 00 33 44

Din analiza exemplului rezultă următoarele:

- Caracterele care nu sunt parte a unei repetări (run) nu suferă nici o prelucrare.
- Run-urile se înlocuiesc cu o tripletă reprezentând caracterul repetat, markerul de repetare 90_H şi lungimea repetării
- În cazul apariției în fluxul de date a markerului 90_H acesta trebuie înlocuit cu o secvență specifică (speculând faptul că run-urile nu au lungime 0)

Chiar dacă în practică putem impune alt mod de reprezentare trebuie să rezolvăm cele două probleme: codarea run-ului şi apariția marker-ului.

Observăm că metoda devine utilă doar în cazul existenței run-urilor de lungime minim 4 (run-uri mai scurte se lasă neschimbate) şi eficiența ei este redusă de existența marker-ului în fluxul de date (caz care duce la o expandare a fluxului).

Aşa cum a fost prezentată metoda se bazează pe modelare **statică** a fluxului de date. Dacă dorim o modelare **semistatică** putem face o analiză a fluxului în avans şi să alegem ca şi marker caracterul care apare de cele mai puține ori în flux. Evident, noul marker trebuie transmis decodorului (de exemplu ca primul caracter din fluxul codat). Dacă dorim o modelare **dinamică** putem face o analiză periodică a fluxului deja transmis şi schimba din mers marker-ul.

4.2. Codare statistică simplă

Pentru exemplificarea acestei metode propunem codarea următorului flux:

ABABACAD

Cea mai simplă codare ar fi codarea pe 2 biți a fiecărui simbol (corespunzătoare unui alfabet cu 4 simboluri) de exemplu astfel:

A – 00, B – 01, C – 10, D – 11

Astfel, secvența codată devine:

0001000100100011

având deci 16 biți.

Codarea statistică presupune realizarea statisticii apariției simbolurilor şi alocarea de coduri mai scurte simbolurilor cu frecvență de apariție mai mare. Evaluăm deci frecvențele de apariție:

A – 4, B – 2, C – 1, D – 1

şi alocăm coduri corespunzător:

A – 0, B – 10, C – 110, D - 111

Fluxul codat devine acum:

01001001100111

având deci doar 14 biți, realizând o compresie față de situația anterioară.

Metoda se încadrează în categoria metodelor **semistatice**, la prima trecere construind modelul (codurile) şi abia la a doua trecere realizând codarea propriu-zisă. Pentru implementarea practică a metodei trebuie transmise decodorului codurile alocate fiecărui simbol, chestiune neglijată anterior, care ar putea compromite câştigul obținut. Deocamdată nu am specificat cum trebuie alese codurile pornind de la frecvențele de apariție (de exemplu prin metodele Shanon-Fano sau mai ales Huffman) dar reținem ideea de a ține cont de statistica fluxului de date (preferabil dezechilibrată).

4.3. Codare bazată pe dicționar şi MTF

Pentru acest exemplu considerăm secvența ABCDDDCCBBAA pe care o codăm bazat pe dicționar. Construim un dicționar cu toate simbolurile (deci care conține ABCD) şi prin codare emitem indexul în acest dicționar. Cazul dicționarului static este prezentat în primele coloane ale tabelul următor:

Dicționar static		
Flux	Dicționar	Index (codare)
A	ABCD	0
B	ABCD	1
C	ABCD	2
D	ABCD	3
D	ABCD	3
D	ABCD	3
C	ABCD	2
C	ABCD	2
B	ABCD	1
B	ABCD	1
A	ABCD	0
A	ABCD	0

Dicționar adaptiv MTF			
Flux	Dicționar curent	Index (codare)	Dicționar actualizat
A	ABCD	0	ABCD
B	ABCD	1	BACD
C	BACD	2	CBAD
D	CBAD	3	DCBA
D	DCBA	0	DCBA
D	DCBA	0	DCBA
C	DCBA	1	CDBA
C	CDBA	0	CDBA
B	CDBA	2	BCDA
B	BCDA	0	BCDA
A	BCDA	3	ABCD
A	ABCD	0	ABCD

Obținem prin codare secvența 012333221100. Evident, deocamdată nu am câştigat absolut nimic.

În continuare considerăm metoda „**M**ove **T**o **F**ront" (MTF) de actualizare a dicționarului. Astfel, după codarea unui caracter pe baza dicționarului acest caracter este „mutat în față". Se încearcă să se favorizeze astfel reprezentarea simbolurilor care se repetă la interval scurt, indexul rezultat fiind în acel caz de valoare mai mică. Metoda este una cu modelare **dinamică** bazată pe dicționar şi, ca toate metodele dinamice, constă în:

- Codare / decodare pe baza modelului curent (dicționarului curent)
- Actualizarea modelului (dicționarului)

În ultimele coloane ale tabelului anterior este prezentată aplicarea metodei MTF pe datele considerate. La fiecare pas dicționarul curent este cel obținut prin actualizarea de la pasul anterior. Fluxul rezultat este 012300102030.

Dacă construim histograma de apariție a indecşilor generați constatăm dezechilibrarea distribuției valorilor în sensul favorizării indecşilor mici:

Fără MTF	Cu MTF
0 ***	0 ******
1 ***	1 **
2 ***	2 **
3 ***	3 **

Acest fapt permite aplicarea asupra fluxului generat a unei metode statistice (de genul celei prezentate anterior) care va avea performanțe bune dacă fluxul este dezechilibrat. Constatăm că metoda este valoroasă ca etapă intermediară într-un lanț mai lung de prelucrări (idee general valabilă în compresia de date). Evident, dacă presupunerea privind repetarea simbolurilor nu este justificată, efectul MTF poate fi nesemnificativ sau chiar negativ. Metoda MTF poate fi folosită şi în cazul în care dicționarele conțin şiruri de simboluri (cazul uzual în practică) nu numai simboluri individuale cum a fost prezentat anterior pentru simplitate. De asemenea se pot lua în considerare şi actualizarea prin mutarea în față doar cu un număr de k poziții (nu până în prima poziție) şi mutarea doar în cazul repetării utilizării unei anumite intrări din dicționar (nu la fiecare utilizare).