ENTROPIA ŞI COMPRESIA DE DATE

Ion Ivan
Daniel Verniş
Doru Ţuţui

Academia de Studii Economice

Bucureşti

Rezumat: Există numeroşi algoritmi de compresie. Performanţa lor depinde de natura informaţiilor din fişiere. Se evaluează entropia fişierelor, se defineşte omogenitatea acestora pentru un set de patru algoritmi de compresie şi se analizează corelaţiile entropie - grad de compresie, după metoda clasică şi cu ajutorul coeficienţilor de corelaţie entropică.

Cuvinte cheie : compresie de date, corelaţie entropică, algoritmi de compresie, entropie.

1. Entropia fişierelor

Fie un fişier iniţial F de date descrise cu alfabetul A, format din simbolurile x_1 x_2.....x_n şi probabilităţile lor de apariţie, respectiv p_1 p_2.....p_n. Câmpul de probabilitate X_F este :

$$X_F \begin{vmatrix} x_1 x_2 \ldots\ldots x_n \\ p_1 p_2 \ldots\ldots p_n \end{vmatrix} \qquad (1)$$

Dacă se consideră fişierele omogene F_1 F_2.....F_m definite cu acelaşi alfabet A şi se concatenează obţinându-se fişierul $FC=F_1 \cup F_2 \cup \ldots\ldots \cup F_m$ rezultă:

$$X_{FC} = \begin{vmatrix} x_1 x_2 \ldots\ldots x_n \\ p_{C1} p_{C2} \ldots\ldots p_{Cn} \end{vmatrix} \qquad (2)$$

probabilităţile p_{C1} p_{C2}.....p_{Cn} au un nivel de obiectivitate mai mare, adică

$$\max\{|p_{ji}-p_{ki}|\} > \max\{|p_{ji} - p_{Ci}|\} \qquad (3)$$

unde p_{ji} este probabilitatea apariţiei şirului x_i în fişierul F_j.

Gradul de nederminare în fişierele F_1 F_2.....F_m depinde de probabilităţile de apariţie a simbolurilor. Această nedeterminare este sesizată prin diferenţa dintre conţinutul fişierelor. Fişierele identice, definite pe un alfabet cu un singur simbol au gradul de nedeterminare minim 0.

Fişiere definite pe alfabete în care simbolurile sunt echiprobabile au gradul de nedeterminare maxim.

Se notează **H(x)** măsura gradului de nederminare egală cu cantitatea medie de informaţie a realizării unui eveniment, dată prin relaţia :

$$H(x) = - \Sigma p_i \log_2(p_i) \qquad (4)$$

Cantitatea de informaţie asociată apariţiei unui simbol x_i este considerată prin convenţie ca având valoarea $-\log_2 p_i$. Se ponderează cantitatea de informaţie cu probabilitatea p_i şi prin însumare rezultă formula unei medii aritmetice:

$$X = \frac{\sum_{i=1}^{n} \alpha_i x_i}{\sum_{i=1}^{n} \alpha_i} \qquad (5)$$

înlocuind α_i cu p_i şi x_i cu $-\log_2(p_i)$ rezultă chiar formula entropiei.

Entropia poate fi considerată o metrică pentru fişierele de date , datorită proprietăţilor:

$$H(X) \geq 0 ; \qquad (6)$$

$$H(X) \leq - \Sigma(1/n)\log_2(1/n) \qquad (7)$$

unde 1/n reprezintă probabilitatea de apariţie a simbolului x_i [n condiţii de echiprobabilitate.

Considerăm două alfabete $A=\{ x_1 x_2 \ldots\ldots x_n \}$ şi $B=\{ y_1 y_2 \ldots\ldots y_p \}$ şi $A \cap B=\phi$. Se construiesc pe cele două alfabete mulţimile de fişiere F_A, F_B.

Fie $F_1 \in F_A$ şi $F_2 \in F_B$ cărora li se asociază câmpurile de probabilitate X_1 respectiv X_2. În aceste condiţii

$$H(X_1,X_2)=H(X_1) + H(X_2) \qquad (8)$$

Dacă alfabetele **A** şi **B** sunt oarecare (nu mai au proprietatea de independenţă)

$$H(X_1,X_2)=H(X_1) + H(X_2/X_1) \qquad (9)$$

unde $H(X_2/X_1)$ este entropia alfabetului sursă **B** condiţionat de apariţia simbolurilor din alfabetul **A**. Deci:

$$H(X_1,X_2) \leq H(X_1) + H(X_2) \qquad (10)$$

$$H(X_1/X_2) = H(X_2/X_1) + H(X_1) - H(X_2) \qquad (11)$$

$$0 \leq H(X_1) - H(X_1/X_2) \leq H(X_1) \qquad (12)$$

Aceste proprietăţi permit efectuarea de cuantificări pe fişiere şi o ierarhizare a acestora în raport cu gradul de nedeterminare.

De exemplu, se consideră fişierele :

$F_1=\{ a a a a b b b b c c c c d d d d \}$

$F_2=\{ a a b b b b c c c c c c c c d d \}$

$F_3=\{ a b b b b c c c c d d d d d d d \}$

în care $lg(F_1) = lg(F_2) = lg(F_3)$, unde lg(.) este funcţia de lungime a unui fişier care indică numărul de simboluri conţinute.

Fişierele F_1, F_2, F_3 sunt definite pe alfabetul $A = \{ a, b, c, d \}$.

Frecvenţele de apariţie ale simbolurilor în cele 3 fişiere sunt date în tabelul 1.

Tabelul 1

Simbol	F1	F2	F3
a	4	2	1
b	4	4	4
c	4	8	4
d	4	2	7

Probabităţile ca frecvenţe relative de apariţie sunt date în tabelul 2:

Tabelul 2

Simbolul	F1	F2	F3
a	1/4	1/8	1/16
b	1/4	1/4	1/4
c	1/4	1/2	1/4
d	1/4	1/8	7/16
Entropia	1.999	1.749	1.771

Prin concatenarea fişierelor F_1, F_2, F_3 rezultă fişierul F_C caracterizat prin valorile date în tabelul 3:

Tabelul 3

Simbolul	Frecvenţa	p_i
a	7	7/48
b	12	12/48
c	16	16/48
d	13	13/48
Entropia	-	1.943

Omogenitatea fişierelor în raport cu distribuţia de probabilitate a simbolurilor poate fi considerată un criteriu de alegere a unui algoritm de compresie de date.

Definind indicatorul normat al entropiei fişierului F definit pe un alfabet cu n simboluri.

$$I_H = \frac{H(F)}{\lg_2 n} \qquad (13)$$

pentru cele 4 fişiere se obţin valorile date în tabelul 4:

Tabelul 4

Fişierul	Entropia	I_H
F1	1.999	0.4997
F2	1.749	0.4372
F3	1.771	0.4427
F4	1.943	0.4857

2. Omogenitatea entropică a fişierelor

Se consideră următoarele loturi de fişiere:

a) **Fişiere de date numerice** care conţin cifre şi separatori sub formă brută, şiruri de caractere numerice, fişiere ce conţin cantităţi, preţuri şi coduri.

Mulţimea este formată din 100 de fişiere ale căror caracteristici sunt date în tabelul 5. Pe ultimele două coloane sunt prezentate valorile entropiei şi ale indicatorului normat, pentru fiecare fişier în parte.

Tabelul 5 - Măsuri ale fişierelor numerice

Fişier	Entropie	L_F	I_H
F1	3.456	2608	0.99950
F2	3.451	1074	0.99823
F3	3.457	2634	0.99974
F4	3.453	944	0.99859
F5	3.451	1750	0.99823
F6	3.442	606	0.99566
F7	3.452	2166	0.99851
F8	3.452	892	0.99829
F9	3.454	2062	0.99909
F10	3.453	1906	0.99872
F11	3.455	2660	0.99931
F12	3.454	1282	0.99887
F13	3.444	866	0.99618
F14	3.452	1698	0.99857
F15	3.455	2556	0.99928
F16	3.450	866	0.99795
F17	3.451	1204	0.99811
F18	3.457	2088	0.99978
F19	3.449	1516	0.99745
F20	3.455	2894	0.99926
F21	3.453	1568	0.99859
F22	3.453	2218	0.99877
F23	3.453	1178	0.99873
F24	3.454	1074	0.99910
F25	3.454	2842	0.99896
F26	3.453	1412	0.99858
F27	3.452	1698	0.99855
F28	3.452	2816	0.99844
F29	3.447	944	0.99692
F30	3.455	1750	0.99921
F31	3.429	450	0.99166
F32	3.453	1750	0.99884
F33	3.448	996	0.99739
F34	3.456	2348	0.99960
F35	3.456	2504	0.99961
F36	3.453	2088	0.99870
F37	3.453	2400	0.99870
F38	3.443	814	0.99582
F39	3.452	684	0.99836
F40	3.446	866	0.99662
F41	3.455	2608	0.99939

F42	3.452	1048	0.99833
F43	3.443	450	0.99577
F44	3.452	1334	0.99840
F45	3.454	2140	0.99902
F46	3.452	944	0.99834
F47	3.452	2842	0.99851
F48	3.454	2868	0.99911
F49	3.452	1360	0.99842
F50	3.448	2504	0.99734
F51	3.452	1776	0.99850
F52	3.457	2660	0.99986
F53	3.455	2582	0.99917
F54	3.449	1802	0.99765
F55	3.455	3024	0.99941
F56	3.455	2114	0.99926
F57	3.456	2348	0.99957
F58	3.453	2582	0.99872
F59	3.453	1464	0.99875
F60	3.443	762	0.99596
F61	3.454	2712	0.99901
F62	3.451	580	0.99821
F63	3.433	476	0.99290
F64	3.452	1100	0.99836
F65	3.452	1854	0.99833
F66	3.455	2114	0.99919
F67	3.455	2504	0.99922
F68	3.451	1412	0.99828
F69	3.447	502	0.99692
F70	3.455	2998	0.99942
F71	3.448	840	0.99725
F72	3.455	1282	0.99925
F73	3.449	1282	0.99769
F74	3.453	1646	0.99871
F75	3.455	2036	0.99918
F76	3.453	2478	0.99883
F77	3.455	1438	0.99918
F78	3.454	2556	0.99899
F79	3.455	2270	0.99918
F80	3.451	1386	0.99815
F81	3.448	918	0.99739
F82	3.451	710	0.99823
F83	3.451	2660	0.99817
F84	3.430	450	0.99214
F85	3.453	1386	0.99866
F86	3.446	580	0.99677
F87	3.447	502	0.99705
F88	3.441	528	0.99527
F89	3.451	1750	0.99823
F90	3.451	918	0.99805
F91	3.455	2816	0.99935
F92	3.448	658	0.99715
F93	3.446	606	0.99682
F94	3.451	1958	0.99827
F95	3.455	2010	0.99934
F96	3.450	502	0.99776
F97	3.457	2842	0.99979
F98	3.455	1542	0.99924
F99	3.456	2530	0.99950
F100	3.452	814	0.99847

b) **Fişiere de text** în care se consideră alfabetul format din totalitatea caracterelor ASCII.

Programul **entropia**.cpp numără frecvenţele de apariţie ale simbolurilor, determină lungimea, calculează entropia şi indicatorul normat .

Rezultatele sunt date în tabelul 6:

Tabelul 6 - Măsuri fişiere text

Fişier	Entropie	L_F	I_H
F101	4.228	654	0.722
F102	4.250	1685	0.699
F103	4.218	991	0.730
F104	4.244	611	0.722
F105	4.979	1052	0.824
F106	3.850	2461	0.635
F107	4.656	564	0.799
F108	4.536	1001	0.778
F109	4.161	1054	0.694
F110	4.975	1580	0.818
F111	4.157	1354	0.693
F112	3.925	3448	0.634
F113	3.972	1462	0.662
F114	4.955	4137	0.793
F115	4.724	431	0.814
F116	4.515	463	0.781
F117	3.694	2733	0.611
F118	4.562	672	0.773
F119	4.171	1832	0.688
F120	4.525	887	0.752
F121	5.023	2480	0.804
F122	5.461	3322	0.872
F123	5.025	1070	0.823
F124	4.982	3334	0.819
F125	4.798	1140	0.823
F126	5.182	1391	0.832
F127	5.178	1390	0.832
F128	5.533	2661	0.866
F129	5.529	4249	0.861
F130	5.461	3322	0.872
F131	5.191	2556	0.842
F132	5.595	10293	0.867
F133	4.954	6252	0.793
F134	5.224	4552	0.844
F135	4.480	1321	0.744
F136	4.812	1606	0.812
F137	4.408	7244	0.741
F138	5.040	5870	0.815
F139	4.850	2146	0.794
F140	4.875	2804	0.781
F141	4.844	5333	0.767
F142	4.859	5805	0.773
F143	4.872	3435	0.798

F144	5.026	3035	0.805
F145	4.361	6830	0.725
F146	4.516	22613	0.694
F147	4.464	19929	0.719
F148	4.738	27141	0.752
F149	4.457	8877	0.735
F150	4.436	24090	0.687

c) **Fişierele de imagine**, pentru care alfabetul este format din totalitatea culorilor şi a nuanţelor de culori folosite într-o imagine. Alfabetul diferă de la o imagine la alta. Rezultatele privind determinarea entropiei şi a indicatorului normat sunt date în tabelul 7.

Tabelul 7 - Măsuri fişiere imagine

Fişier	L_F	Entropie	I_H
F151	60406	2.286	0.481
F152	129078	5.713	0.716
F153	32278	7.424	0.928
F154	38462	6.221	0.784
F155	69167	6.321	0.796
F156	33478	6.898	0.868
F157	35478	5.599	0.717
F158	140918	5.499	0.761
F159	308278	5.723	0.724
F160	308278	5.746	0.732
F161	100118	1.217	0.234
F162	248038	4.755	0.607
F163	157042	5.892	0.741
F164	150	2.384	0.596
F165	838	1.226	0.222
F166	838	2.243	0.392
F167	838	1.265	0.236
F168	5318	3.111	0.593
F169	630	2.255	0.522
F170	630	2.129	0.493
F171	20710	3.153	0.589
F172	20710	3.117	0.565
F173	2710	2.905	0.462
F174	2710	1.819	0.375
F175	19258	1.840	0.341
F176	19258	2.085	0.307
F177	2710	2.281	0.357
F178	2710	2.228	0.480
F179	2710	0.670	0.168
F180	630	2.292	0.540
F181	630	2.162	0.509
F182	2710	0.618	0.148
F183	350	2.288	0.506
F184	838	1.259	0.228
F185	790	2.907	0.558
F186	790	2.724	0.567
F187	838	1.646	0.293
F188	838	1.776	0.309
F189	307514	4.991	0.767
F190	66146	6.283	0.837
F191	350	2.618	0.507
F192	838	2.100	0.365
F193	838	2.092	0.373
F194	838	1.673	0.308
F195	838	2.170	0.391
F196	838	3.072	0.529
F197	157042	5.892	0.741
F198	838	1.732	0.314
F199	339178	3.438	0.454
F200	161078	4.407	0.617

d) **Fişierele de sunet**, pentru care alfabetul este format din totalitatea tonurilor şi a notelor muzicale folosite într-o melodie. Alfabetul diferă de la un fişier la altul. Rezultatele privind determinarea entropiei şi a indicatorului normat sunt date în tabelul 8.

Tabelul 8 - Măsuri fişiere sunet

Fişier	L_F	Entropie	I_H
F201	1874	6.528	0.858
F202	4460	5.346	0.712
F203	15932	5.247	0.759
F204	2624	6.323	0.849
F205	8548	6.656	0.849
F206	110316	5.916	0.766
F207	6596	6.045	0.811
F208	3202	6.261	0.834
F209	5694	6.055	0.788
F210	4400	5.793	0.767
F211	4456	6.381	0.851
F212	9450	6.193	0.796
F213	10328	5.572	0.723
F214	3740	6.323	0.824
F215	45662	6.501	0.815
F216	5982	6.441	0.832
F217	27804	5.840	0.823
F218	11586	4.822	0.690
F219	7676	6.305	0.819
F220	4772	5.938	0.800
F221	8622	6.307	0.803
F222	15082	6.590	0.834
F223	6954	6.167	0.788
F224	3278	5.508	0.733
F225	3638	5.594	0.764
F226	6964	6.182	0.788
F227	10374	6.481	0.825
F228	10344	6.099	0.797
F229	6100	6.320	0.804
F230	6242	6.321	0.806
F231	89126	6.215	0.777
F232	143914	6.895	0.862
F233	175146	6.927	0.866
F234	140330	6.717	0.840
F235	166954	6.996	0.875
F236	147754	7.042	0.881
F237	169010	6.584	0.823
F238	74026	5.763	0.721
F239	142888	6.110	0.764

F230	169010	6.797	0.850
F241	145446	6.929	0.867
F242	159782	6.873	0.860
F243	145450	6.348	0.794
F244	184872	6.696	0.837
F245	129578	6.367	0.796
F246	26272	6.216	0.795
F247	13350	6.604	0.842
F248	275516	3.669	0.471
F249	3464	6.432	0.836
F250	24994	3.835	0.536

Se construieşte o listă de perechi (I_{Hi}, α_j) j=1...M unde:

M - are valoarea de 250, reprezentând numărul total de fişiere;

I_{Hi} - reprezintă indicatorul normat al entropiei fişierului cu numărul curent j;

α_j - reprezintă tipul fişierului care poate fi:

n - fişier numeric dacă aparţine mulţimii fişierelor date în tabelul 5;

t - fişier text dacă aparţine mulţimii fişierelor date în tabelul 6;

i - fişier imagine dacă aparţine mulţimii fişierelor date în tabelul 7;

s - fişier sunet dacă aparţine mulţimii multimii fişierelor date în tabelul 8.

Se ordonează perechile după nivelul entropiei, lista împărţindu-se în 4 subliste după criteriul omogenităţii entropice.

Se obţin următoarele rezultate:

Tabelul 9

Grup	*n*	*t*	*i*	*s*
Grup 1	0	0	0.024	0
Grup 2	0	0	0.068	0.004
Grup 3	0	0.072	**0.080**	0.024
Grup 4	**0.400**	**0.128**	0.028	**0.172**

unde la intersecţia liniei i cu coloana j se află ponderea fişierului de tip j aflat în grupa i. .

Asupra fiecărui fişier din grupurile create se aplică următorii algoritmi de compresie :

- Algoritmul Huffman standard

- Compresia aritmetică

-Algoritmul de precompresie RLE (Run Length Encoding)

- Algoritmul LZW (Lempel Ziv Welch)

Pentru primul grup de fişiere, rata de compresie cea mai mare este obţinută cu algoritmul RLE. După cum se observă, în grupul unu de fişiere sunt incluse doar fişiere imagine. Principalele caracteristici ale acestor fişiere de imagini sunt :

- lungimea destul de mare în comparaţie cu celelalte tipuri de fişiere ;

- un alfabet redus de simboluri ;

- succesiuni de simboluri identice de lungime mare.

Aceste caracteristici explică şi rezultatul bun obţinut cu algoritmul RLE.

Pentru cel de-al doilea grup de fişiere, algoritmul Huffman, urmat de compresia aritmetică dau cele mai bune rezultate. Fişierele din acest grup mai păstrează încă unele din caracteristicile grupului 1. Caracteristicile predominante în acest caz:

- lungimea destul de mare în comparaţie cu celelalte tipuri de fişiere ;

- un alfabet redus de simboluri.

Algoritmii Huffman şi compresia aritmetică dau rezultate destul de bune în cazul unor alfabete de lungimi reduse.

Pentru al treilea grup, ratele de compresie cele mai bune s-au obţinut cu algoritmul de compresie aritmetică. Caracteristicile evidenţiate pentru acest grup:

- n lungimea medie a fişierelor;

- un alfabet destul de bogat în simboluri.

Grupul al patrulea este grupul fişierelor cel mai greu de compresat. Pentru majoritatea din fişierele incluse în acest grup algoritmii RLE şi compresia aritmetică nu au putut obţine rate de compresie bune. Singurul algoritm care a reuşit rate de compresie convenabile pentru acest grup este algoritmul LZW. Rata de compresie a fost în medie de 30-40% din lungimea fişierului iniţial.

De reţinut faptul că algoritmul LZW este specific compresiei fişierelor de lungimi mari, însă în cazul grupului 1, succesiunea de simboluri identice a fost caracteristica predominantă.

3. Corelaţia entropie - grad de compresie

Se consideră mulţimea formată din fişierele **F1, F3, F29, F33, F45, F56, F62, F77, F85, F91, F110, F114, F124, F132, F148, F151, F165, F175, F185, F195, F201, F215, F225, F235, F245**, selectate aleator din mulţimea fişierelor analizate.

Pentru aceste fişiere se realizează compresia folosind algoritmii Huffman, Compresia aritmetică (CA), LZW, RLE [7], [8]. Datele rezultate sunt prezentate în tabelul 10.

Se calculează coeficienţii de corelaţie [ntre variabilele lungime, entropie, indicatorul normat al

entropiei]i gradele de compresie ai algoritmilor obţin[ndu-se rezultatele din tabelul 11.

între rezultatele acestui algoritm şi indicatorul entropic. Algoritmul RLE are la bază principiul

Tabelul 10 - Caracteristici de compresie/entropie

Fişier	Entropie	Lungime	I_H	Huffman	CA	LZW	RLE
F 1	3.456	2608	0.99950	50.87	63.69	42.51	100
F 3	3.457	2634	0.99974	50.05	64.04	42.57	99.97
F 29	3.447	944	0.99692	49.15	64.76	43.62	100
F 33	3.448	996	0.99739	51.97	65.42	43.82	99.82
F 45	3.452	2140	0.99902	52.66	64.61	42.69	99.83
F 56	3.455	2114	0.99926	50.48	64.20	42.77	99.96
F 62	3.451	580	0.99821	47.12	66.98	45.71	100
F 77	3.455	1438	0.99918	48.07	65.14	43.32	99.81
F 85	3.453	1386	0.99866	49.82	65.81	43.61	100
F 91	3.455	2816	0.99935	54.31	64.05	42.44	99.97
F 110	4.975	1580	0.818	75.95	46	80.00	98.16
F 114	4.955	4137	0.793	67.97	62	71.79	85.67
F 124	4.982	3334	0.819	68.93	73	65.39	92.83
F 132	5.595	10293	0.867	73.06	72	67.44	99.32
F 148	4.738	27141	0.752	57.65	70	71.91	100.0
F 151	2.286	60406	0.481	28.91	94	15.34	21.12
F 165	1.226	838	0.222	37.83	77	25.30	36.63
F 175	1.840	19258	0.341	25.44	93	15.70	23.83
F 185	2.907	790	0.558	51.90	67	43.80	82.41
F 195	2.170	838	0.391	46.06	70	35.32	54.42
F 201	6.528	1874	0.858	87.90	6	95.71	99.95
F 215	6.501	45662	0.815	83.41	69	95.03	99.83
F 225	5.594	3638	0.764	83.75	34	74.71	90.19
F 235	6.996	166954	0.875	64.01	17	92.89	98.91
F 245	6.367	129578	0.796	72.03	74	97.62	98.56

Tabelul 11 - Coeficienţi de corelaţie

	H	L	I_H	G_{Huff}	G_{CA}	G_{LZW}	G_{RLE}
H	1	0.4949	0.4041	0.8926	-0.6322	0.9641	0.6042
L	0.4949	1	-0.0698	0.1386	-0.1727	0.4509	-0.0218
I_H	0.4041	-0.0698	1	0.3384	-0.3465	0.3017	0.8830
G_{Huff}	0.8926	0.1386	0.3384	1	0.6516	0.9180	0.6126
G_{CA}	-0.6322	-0.1727	-0.3465	0.6516	1	-0.6386	-0.4981
G_{LZW}	0.9641	0.4509	0.3017	0.9180	-0.6386	1	0.5875
G_{RLE}	0.6042	-0.0218	08830	0.6126	-0.4981	0.5875	1

Din analiza tabelului 11 se observă că cele mai puternice legături între două variabile sunt cele dintre algoritmul de compresie LZW şi entropia fi]ierelor, respectiv algoritmul Huffman şi entropia fi]ierelor. Aceste rezultate sunt în concordanţă cu valorile din tabelul 9, unde ponderea cea mai mare a fi]ierelor se află în grupul 4, grup caracteristic algoritmului LZW.

O altă legatură puternică se observă şi între rezultatele algoritmilor Huffman şi LZW, aceasta explic[nd faptul că cele două metode de compresie au aplicabilitate la majoritatea fişierelor. În cazul algoritmului RLE, se observă o legatură puternică reducerii redundanţei într-un fişier, aceasta fiind specifică fişierelor imagine. Indicatorul entropic arată exact această redundanţă a fişierelor. Elementele de neomogenitate a comportamentului algoritmului sunt date în [9].

4. Corelaţia entropică şi compresia de date

Se prezintă, în continuare, principalii indicatori entropici, care se folosesc la studiul fişierelor

prezentate. Simbolurile folosite au semnificațiile din [1] și [2]:

X - tipul fișierului (text, numeric, imagine, sunet);

Y - lungimea fișierului;

Z - rata de compresie.

WXYZ = H(X)+H(Y)+H(Z) - H(X,Y,Z) corelația entropică globală a variabilelor X, Y, Z;

WXZCY = H(Y)-H(Y/(X,Z)) corelația entropică a vectorului (X,Z) și a variabilei Y;

WYZCX = H(X)-H(X/(Y,Z)) corelația entropică a vectorului (Y,Z) și a variabilei X;

Tabelul 12

Indicatorul entropic	Algoritmul Huffman	Compresia aritmetică	Algoritmul LZW	Algoritmul RLE
$H_{(Z/X)}$	2.407111	2.357246	1.920330	1.562063
$H_{(Z/Y)}$	2.275095	2.346348	2.157537	1.561923
W_{XZ}	0.531875	0.748993	0.949253	0.454620
W_{YZ}	0.663891	0.759891	0.712046	0.454760
$H_{(XYZ)}$	5.065323	5.230175	4.842816	4.721472
$H_{((XZ)/Y)}$	2.529880	2.694732	2.307373	2.186029
H((YZ)/X)	3.579156	3.744008	3.356648	3.235305
H(Z/(XY))	1.427742	1.592593	1.205253	1.083890
WXYZ	1.895274	1.897674	2.048377	1.316822
WXZCY	1.363399	1.148681	1.099123	0.862201
WYZCX	1.231382	1.137783	1.336330	0.862061
RXYZ	0.187083	0.181416	0.211486	0.139450
RXZCY	0.269163	0.219626	0.226960	0.182613
RYZCX	0.243100	0.217542	0.275941	0.182583

Relațiile folosite pentru corelația entropică sunt:

WXZ = H(X)+H(Y)-H(X,Y) corelația entropică globală a variabilelor X, Z;

WYZ = H(Y)+H(Z)-H(Y,Z) corelația entropică globală a variabilelor Y, Z;

$$H(Z/X) = -\sum_{j=1}^{m}\sum_{i=1}^{n} p(z,x)\cdot \log p(z/x)$$

entropia variabilei Z condiționată de variabila X;

$$H(Z/Y) = -\sum_{j=1}^{m}\sum_{k=1}^{p} p(z,y)\cdot \log p(z/y)$$

entropia variabilei Z condiționată de variabila Y;

$$H(XYZ) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{p} p(x_i,y_j,z_k)\cdot \log p(x_i,y_j,z_k)$$

entropia globală a variabilelor X,Y,Z;

H((XZ)/Y) = H(X,Y,Z) - H(Y) entropia;

RXYZ =WXYZ/(2 · H(X,Y,Z)) coeficient de corelație entropică globală a variabilelor X,Y, Z;

RXZCY = WXZCY/H(X,Y,Z) coeficient de corelație entropică a vectorului (X,Z) și a variabilei Y;

RYZCX = WYZCX/H(X,Y,Z) coeficient de corelație entropică a vectorului (Y,Z) și a variabilei X.

Indicatorii prezentați sunt diferiți de zero. Deci, în toate cazurile, există interdependențe atât între componentele vectorului, gândite ca un ansamblu (WXYZ), cât și între perechi de componente și respectiv cea de a treia componentă.

Se observă că cele mai puternice interdependențe ale celor trei variabile se înregistrează în cazul algoritmului LZW, iar cele mai slabe în cazul algoritmului RLE. Totuși, valoarea coeficienților de corelație entropică globală în cazul algoritmilor Huffman și Compresia aritmetică sunt foarte apropiate, și nu la mare diferență față de coeficientul similar în cazul algoritmului LZW.

Valoarea coeficientului RXZCY este cea mai mare în cazul algoritmului Huffman, ceea ce ne

indică faptul că rezultatele pentru algoritmul Huffman împreună cu tipul de fişier sunt cele mai influenţabile de către lungimea fişierului. Valoarea coeficientului R_{YZCX} ne indică faptul că în cazul algoritmului LZW avem cea mai mare corelaţie între perechea de variabile rata de compresie - lungimea fişierului şi variabila tip fişier.

Analizând valorile entropiilor şi ale corelaţiilor entropice dintre ratele de compresie şi tipul fişierului, respectiv lungimea fişierului, vom găsi:

- entropia ratei de compresie, condiţionată de tipul fişierului este cea mai mare în cazul algoritmului Huffman;

- entropia ratei de compresie condiţionată de lungimea fişierului este cea mai mare în cazul compresiei aritmetice.

Concluzii

În literatura de specialitate recentă, entropia este folosită drept criteriu de neomogenitate ale fişierelor. Analiza entropică permite stabilirea apartenenţei la o anumită clasă de entropie şi în acest fel se va asocia direct componenta program, cores-punzătoare algoritmului care care realizează compresie performantă. Se fundamentează construirea de software pentru compresie care ţine seama de neomogenitatea fişierelor.

Bibliografie

1. **PURCARU, I.**: Informaţie şi corelaţie , Editura Ştiinţifică si Enciclopedică , Bucureşti, 1988.
2. **TOVISSI, PURCARU, I., IVAN, I.**: Corelaţia entropică pentru trei caracteristici ale unei colectivităţi. În: Revista de statistică, nr. 2, 3, 1981.
3. **DODESCU, GH., IONESCU, D.**: Sisteme electronice de calcul şi teleprelucrare, Editura Didactică şi Pedagogică, 1980.
4. **IVAN, I., VERNIŞ, D.**: Analiza comparată a algoritmilor de compresie date. În: PC World, nr.12, decembrie 1995.
5. **IVAN, I., VERNIŞ, D.**: Evaluarea seturilor de date destinate compresării. În: Revista de statistică, nr.1, ianuarie 1996.
6. **IVAN, I., VERNIŞ, D.**: Compresia de date. În: PC Report, nr.9, septembrie, 1996.
7. **VERNIŞ, D., I. IVAN, P. OPREA**: Metrici ale fişierelorpentru compresie. În: Revista de Informatică Economică, nr.2, 1997.
8. **PLUME, P.**: Compression du donnees, Editura Eyrolles, Paris, 1993.
9. **VERNIŞ, D.**: Neomogenitatea fişierelor în compresia de date. Sesiunea ştiinţifică, ASE, mai 1995.