

Algoritm pentru compresia fisierelor PHBO

Prof.dr. Ion IVAN

Catedra de Informatica Economica, A.S.E. Bucuresti

Prog. Daniel VERNIS

Pagini Aurii România, Bucuresti

În momentul de fata exista o multitudine de algoritmi de compresie pe piata internationala de software. Însa, algoritmul VI prezentat în aceasta lucrare încearca o noua abordare a compresiei de date, în sensul prelucrării fisierelor în cascada, cu o serie de algoritmi performanti.

Cuvinte cheie: compresie de date, BWT, VI, PHBO, sortare.

1. Introducere

Societatea informationala este caracterizata prin manipularea unor baze de date foarte mari. Baza de date a firmelor înregistrate la Registrul Comertului, Mersul Trenurilor, reprezentarea localitatilor si a obiectivelor turistice, Nomenclatorul de produse si multe alte fisiere stau la baza unor aplicatii actuale de comert electronic, informare curenta sau pentru fundamentarea deciziilor. Chiar si în conditiile dezvoltarii hardware, în care capacitatea de memorare pe HDD depaseste ordinul de marime GIGA, resursa memorie ramâne restrictiva datorita noilor probleme ce se formuleaza odata cu evolutiile din planul software. Compresia de date este o solutie de eliminare a restrictiilor impuse de capacitatea de stocare. În [1] s-a evidentiat faptul din zece algoritmi si cincisprezece tipuri de fisiere, unii algoritmi ofera performante mai bune pentru anumite tipuri de fisiere. În [2] s-au prezentat o serie de criterii pentru alegerea algoritmului optim de compresie pentru un anumit tip de fisiere, dupa efectuarea unei analize pe text.

Articolul de fata porneste de la necesitatea construirii unui algoritm specializat pentru compresie / decompresie care sa scada cât mai mult posibil gradul de compresie r .

Exemplificarea se va face pe un fisier al abonatilor telefonici. Problema presupune un numar mare de posturi telefonice si instrumente adecvate de regasire a numerelor de telefon dupa diferite criterii (nume, adresa, starda, sau combinatii ale acestora).

2. Particularitati ale fisierelor PHBO

Pentru a construi un algoritm specializat, se efectueaza o analiza a fisierului sau a clasei de fisiere. Rata de compresie este cu atât mai buna cu cât algoritmul este mai particular, specific pentru o anumita tipologie de fisiere si în consecinta este aplicat acestei clase de fisiere (vezi JPEG).

Fisierele PHBO sunt caracterizate prin:

- Structura fixa de articol (în cazul de fata: nume [L1], adresa [L2] si numar de telefon [L3]);
- Numar de articole [N], valoare de ordinul miilor si sutelor de mii;
- Pentru regasire se efectueaza sortare alfabetica multicriteriala;
- Se înregistreaza repetitivitatea la nivelul numelor, adreselor si numerelor de telefon;
- Exista un nomenclator limitat al strazilor;
- Numerele de telefon cu 6-7 cifre reprezinta o plaja acoperitoare care determina particularitati de structurare pentru L1 si L2. Toate acestea indica un grad de redundanta ridicat în fisierele PHBO. Pentru a masura redundanta se calculeaza indicatorul I_{STS} , care reprezinta ponderarea diferentelor înregistrate fata de elementul precedent, la nivel de sir. De exemplu, se considera sirurile:

Tabelul 1. Modul de calcul al I_{STS}

Sir	Cod sir	Indicator I_{STS}
aabb	(s1)	0
aabb	(s2)	0
acdd	(s3)	1

axab	(s4)	2
baaa	(s5)	3
baaa	(s6)	3
baaa	(s7)	3
cdab	(s8)	4
cdac	(s9)	5

Indicatorul I_{STS} arata câte dintre sirurile vecine (adiacente) sunt diferite. La aparitia unui sir s_i , diferit de sirul precedent s_{i-1} , I_{STS} se incrementeaza cu o unitate. Folosind indicele I_{STS} , se calculeaza I_{ST} dupa urmatoarea relatie: $I_{ST} = (I_{STS} / N) * 100$.

Asupra fisierului PHBO se realizeaza o sortare multicriteriala si se calculeaza trei indici I_{ST} , câte unul pentru fiecare grup. În cazul abordat, vom avea I_{STnume} , $I_{STadresa}$ si $I_{STtelefon}$. Algoritmul de compresie propus în continuare va opera asupra fisierelor PHBO, folosind valorile precalculate ale indicatorilor I_{ST*} .

3. Etapele compresiei fisierelor PHBO

Algoritmul pentru compresia fisierelor PHBO este o varianta simplificata a algoritmului VI. Pentru a realiza o compresie eficienta asupra fisierelor de tip PHBO se vor parcurge urmatoorii pasi:

1. Se delimiteaza coloanele corespunzând câmpurilor prezentate anterior, **L1**, **L2** si **L3**;
2. Se calculeaza I_{ST*} pentru cele trei câmpuri;
3. Pentru fiecare coloana în parte se executa urmatoarele operatii:
 - 3.1. Se aplica transformarea BWT de grad I, astfel:
 - 3.2. Se determina doua variabile: **S** de tip string si **Nn**, care reprezinta lungimea sirului;
 - 3.3. Se genereaza **Nn** siruri virtuale, de lungime **Nn**, dupa urmatul algoritim:

S_i va fi format din $S_{i-1}[2..Nn] + S_{i-1}[1]$, unde **[2..Nn]** reprezinta caracterele din sir de pe pozitia 2 pâna la pozitia Nn, iar **[1]** caracterul de pe prima pozitie din sirul S_{i-1} .

Observatie: Notiunea de sir care este folosita în acest algoritim nu este aceeași cu

aceea din limbajele C/C++ sau Pascal. În acest algoritim sirul este o însiruire de octeti. Un sir virtual este un sir care remapeaza zona de memorie a sirului initial, în cazul nostru **S**.

De exemplu, pentru sirul POPESCU vor fi generate urmatoarele siruri virtuale:

s_0	P	O	P	E	S	C	U
s_1	O	P	E	S	C	U	P
s_2	P	E	S	C	U	P	O
s_3	E	S	C	U	P	O	P
s_4	S	C	U	P	O	P	E
s_5	C	U	P	O	P	E	S
s_6	U	P	O	P	E	S	C

3.4. Se realizeaza o sortare lexicografica a sirurilor $s_0..s_6$. În cazul sirului prezentat va rezulta urmatoarea lista de siruri sortate:

s_5	C	U	P	O	P	E	S
s_3	E	S	C	U	P	O	P
s_1	O	P	E	S	C	U	P
s_2	P	E	S	C	U	P	O
s_0	P	O	P	E	S	C	U
s_4	S	C	U	P	O	P	E
s_6	U	P	O	P	E	S	C

3.5. Dupa sortare, se preiau doar prima si ultima coloana din tabela de siruri (lista). Pentru exemplul prezentat vom avea coloanele **F** si **L**:

	F						L
s_5	C	U	P	O	P	E	S
s_3	E	S	C	U	P	O	P
s_1	O	P	E	S	C	U	P
s_2	P	E	S	C	U	P	O
s_0	P	O	P	E	S	C	U
s_4	S	C	U	P	O	P	E
s_6	U	P	O	P	E	S	C

F: CEOPPSU

L: SPPOUEC

3.6. Caracterele din coloana **L** nu par a avea o ordine determinata, dar de fapt, aceasta coloana are o proprietate speciala. Fiecare

caracter din coloana **L**, este un *caracter prefix* pentru sirul care porneste din acelasi rând dar din coloana **F** [vezi 1].

3.7. Pentru procesul de compresie, se va pastra doar coloana **L** si un *index de start* care va puncta spre rândul care contine sirul initial. În cazul prezentat, *indexul de start* va fi: 2.

Indexul de start este egal cu 2, valoare care reprezinta rândul pe care se gaseste, în coloana **L**, caracterul cu care începe sirul **S**, în cazul discutat '**P**'.

Se memoreaza în fisierul compresat doar *indexul de start* si valorile din coloana **L**, la decompresie, restul informatiei regenerându-se din acestea.

3.9. Pentru ca algoritmul sa fie mai rapid, lungimea sirului **S** trebuie sa nu fie foarte mare. Prin teste repetate, s-a ajuns la concluzia ca pentru fisierele PHBO, o lun-

gime de **6000-8000** de caractere genereaza cele mai bune performante.

3.10. În header-ul fiecarui fisier compresat (în cazul prezentat avem 3 fisiere intermediare), vom avea urmatoarele valori: lungimea normala a unui buffer si lungimea ultimului buffer din fisier.

4. Se concateneaza toate fisierele compresate intermediare, obținându-se un fisier compact **F_T**.

5. Asupa fisierului **F_T** se aplica o transformare MTF. Prin această algoritm se uniformizeaza transformarile din **F_T**, pentru a se putea aplica un algoritm de compresie fara pierdere de informatie.

4. Rezultate experimentale

Se considera fisierele de tip PHBO, corespunzatoare unor judete ale caror continut este prezentat partial în Tabelul 2.

Tabelul 2. Continutul fisielerelor PHBO

Fisier	Lungime	Numar de articole	I _{ST} *
Ialomita.phbo	12.152.112 octeti	29.424	62,4244%
Constanta.phbo	55.220.578 octeti	133.706	65,6410%
Giurgiu.phbo	8.866.697 octeti	21.469	68,3776%
Dimbovita.phbo	18.871.209 octeti	45.693	65,2901%
Vrancea.phbo	16.060.744 octeti	38.888	67,0901%
Arad.phbo	27.089.909 octeti	65.593	61,3511%

Rularea aplicatiei care a implementat algoritmul a condus la urmatoarele rezultate ale compresiei (tabelul 3):

Tabelul 3. Rezultate experimentale

Fisier	Lungime initiala	WinZIP	WinRar	Transformari BWT/MTF	VI _H	VI _{LZW}
Ialomita.phbo	12.152.112	568.982	488.486	12.153.636	432.426	441.518
Constanta.phbo	55.220.578	3.230.860	2.628.393	55.227.485	2.390.836	2.102.486
Giurgiu.phbo	8.866.697	424.423	359.910	8.867.810	324.792	290.662
Dimbovita.phbo	18.871.209	928.385	761.902	18.873.572	760.580	762.190
Vrancea.phbo	16.060.744	904.966	724.091	16.062.756	690.322	612.140
Arad.phbo	27.089.909	1.628.717	1.352.063	27.093.300	1.350.458	1.224.164

Din Tabelul 3 rezulta oportunitatea efectuării compresiei cu algoritmul VI. Algoritmul **VI_{PHBO}** se dovedeste a fi superior pentru cazul particular de structura a fisielerelor PHBO. Algoritmul **VI_{PHBO}** folosit pentru un alt tip de structura, dar cu indicatorul

I_{ST} recalculat, poate da rezultate la fel de bune, fara modificarea codului aplicatiei.

Concluzii

Existenta unei colectii de algoritmi specifici pentru anumite tipuri de fisiere conduce la

generarea de aplicatii performante, care vor avea rezultate foarte bune în masura în care vor exista mai multe tipuri de fisiere tratate. În cadrul unei aplicatii care trateaza distinct tipuri de fisiere diferite, pentru a obtine mereu rezultate performante va trebui inclus si un algoritm de determinare a performantelor algoritmilor de compresie, în cazul în care se întâlnește un nou tip de fisiere. Aceste aspecte ale compresiei de date folosind ierarhii si istorii de tipuri de fisiere au fost tratate pe larg în [1].

Bibliografie

1. COMPRESIA DE DATE, Ion Ivan, Daniel Vernis, CISON 1998
2. EVALUAREA SETURILOR DE DATE DESTINATE COMPRESARII, Ion Ivan, Daniel Vernis, Revista Româna de statistica nr 1 / 1996
3. ENTROPIA SI COMPRESIA DE DATE, Ion Ivan, Daniel Vernis, Doru Tutui, Buletinul de informatica, ICI nr 3 / 1997

Comisia 11 a Colegiului Consultativ pentru Cercetare, Dezvoltare si Inovare anunta Organizarea competitiei publice pentru teme noi de Cercetare-Dezvoltare si etape noi la teme în derulare, pe anul 1999, încadrate în programele componente si obiective specifice *Informatica*, din cadrul directiei tematice corespunzatoare a Programului *Orizont 2000* (aprobat prin HG nr.1095/1995 si publicat în Monitorul Oficial al României din 10 ianuarie 1996). Propunerile de teme noi, pentru care nu se preved încheierea în anul 1999, vor fi corelate cu obiectivele programului **Tehnologii pentru Societatea Informationala** din structura Planului National de Cercetare-Dezvoltare si Inovare, asa cum a fost aceasta facuta publica, ca proiect, în materialele difuzate la *Simpozion Dezvoltare prin Inovare si Parteneriat*, din 19. XI. 1998 si care se afla la dispozitia celor interesati la sediul Comisiei. Informatii suplimentare, lista obiectivelor scoase la competitie si documentatia pentru ofertare pot fi obtinute de la Secretariatul Comisiei, ASE Bucuresti (Prof.dr. I.Gh. Rosca), Departamentul de Cercetari Economice, str. Calea Dorobanti nr. 15-17, sector 1, camera 403C, Bucuresti, tel/fax 2116450. Depunerea ofertelor se face pâna la 28.12.1998, ora 16,³⁰

Anunt publicat în ziarul "Adevarul" din 15 decembrie 1998