

Semestrální práce z předmětu 36PJC

Martin Tůma, 3/50

7.6.2005

Obsah

1	Zadání	3
1.1	Popis programu	3
1.2	Použití	3
2	Implementace	3
2.1	Popis řešení	3
2.2	Implementace	3
3	Závěr	4

1 Zadání

1.1 Popis programu

tzip je jednoduchý archivační program používající kompresní algoritmus LZW. Program umožňuje kromě komprese jednotlivých souborů také archivaci a následnou kompresi adresářů při zachování jejich struktury. Předlohou pro program byl známý archivační program *tar*, *tzip* tak implementuje obdobu `tar -cvvzf` pro kompresi a `tar -xvvzf` pro dekompresi.

1.2 Použití

```
tzip -x|c archiv (soubor1|adresář1 ... [souborN|adresářN])
  -c zkomprimuje soubory|adresáře do archivu
  -x extrahuje archiv
```

2 Implementace

2.1 Popis řešení

LZW je speciálním druhem „slovníkové“ komprese. Funguje tedy tak, že opakující se řetězce zapisuje jako číselné kódy. Výhodou LZW ale je, že slovník není součástí archivu, neboť je při dekompresi postupně vytvářen z již načtených dat. Podrobnější popis algoritmu viz [1].

Pro archivaci souborů a adresářů (tzn. zapsání více souborů/adresářů do jednoho souboru) je použit formát podobný formátu *tar*. Soubory jsou v archivu zaznamenány za sebou tak, že na začátku každého souboru v archivu je hlavička obsahující informace (jméno, práva, velikost...) o následujícím souboru.

Program pak pracuje tak, že načítá soubory, ty „formátuje“ a předává je dále kompresnímu algoritmu, který je komprimuje. Protože při LZW kompresi není nutné znát předem všechna data určená ke komprimaci, mohou být data komprimována „on the fly“ již během načítání souborů.

2.2 Implementace

Program je rozdělen na 3 moduly. Modul *ttr* zajišťuje archivaci souborů, modul *lzw* komprimaci a hlavní program *tzip* pak jejich propojení a ovládání programu.

LZW - algoritmy použité v komprimátoru jsou lehce modifikovanou verzí uvedenou v [1]. V programu je použito 16b kódování, tzn. slovník může obsahovat maximálně $2^{16} - 256$ řetězců. Vyhledávání ve slovníku je implementováno jednoduchou hashovací funkcí. Knihovna implementuje tyto 2 funkce:

```
/*
** lzw_compress - zkomprimuje vstupní proud "input" do souboru "output"
```

```
** lzw_extract - dekomprimuje vstupní soubor "input" do proudu "output"  
*/
```

```
extern int lzw_compress(int input, FILE *output);  
extern int lzw_extract(FILE *input, int output);
```

TTR - Knihovna pro archivaci. Implementuje tyto 2 funkce:

```
/*  
** ttr_add_file - přidá soubor "file_name" do proudu "fd"  
** ttr_extract_file - extrahuje soubor z proudu "fd"  
*/
```

```
extern int ttr_add_file(int fd, char *file_name);  
extern int ttr_extract_file(int fd);
```

Hlavní program pak vytvoří pro archivátor a komprimátor samostatné procesy a ty navzájem pomocí roury propojí.

3 Závěr

Při testech účinnosti komprimátoru bylo dosaženo poměrně slušných výsledků a to jak na textových, tak na binárních datech. Například při kompresi adresáře */etc* to je 65%, u adresáře */usr/bin* pak 75%. Nejlepších výsledků program dosahuje na souborech XLS, či DOC, kde může komprese dosáhnout až 15%! Naopak pokud vstupní data mají výrazně nehomogenní charakter, jsou výsledky velice slabé a při pokusu o aplikování programu na již komprimovaná data, může výstup oproti vstupu i narůst (někdy i o 30-40%).

Protože aplikace využívá některé UNIX specifické funkce, je její přenositelnost omezena na POSIX kompatibilní systémy. Zaručeně lze přeložit (testováno) na těchto OS: Linux 2.6.x, Solaris 9, FreeBSD 4.x.

Přestože byl program důkladně testován a při jeho testování jsem nenarazil na žádné „neočekávané události“, je jeho použití jen na vlastní nebezpečí a autor nenese žádnou zodpovědnost za programem způsobené škody!

Použitá literatura

- [1] *LZW Data Compression*
<http://www.dogma.net/markn/articles/lzw/lzw.htm>
- [2] *GNU TAR*
<http://www.gnu.org/software/tar/tar.html>

- [3] *The GNU C Library*
http://www.gnu.org/software/libc/manual/html_node/
- [4] *Materiály k 36OSY*
<http://service.felk.cvut.cz/courses/X36OSY/>