

Kompresia údajov.

Pri prevode dát (najmä obrázky a zvuky) z analógovej reprezentácie do digitálnej vznikajú problémy s veľkosťou súborov a práca s takýmito súbormi je náročná a pomalá. Pri pripájaní prílohy emailu, ktorá sa skladá z viacerých súborov, vzniká problém s načítavaním a pripájaním každého z nich, pričom treba dávať pozor na to, aby sme nepresiahli maximálnu povolenú veľkosť prílohy. Preto sa vyvinuli metódy, ako už zapísané údaje znovu prekódovať – komprimovať (komprimovať, zhustiť, pakovať) tak, aby zaberali čo najmenší objem dát na disku alebo vytvorili jeden „nedeliteľný“ priečinok.

Kompresia údajov je séria metód ako prekódovať údaje tak, aby zaberali menší objem pri zachovaní pôvodného obsahu (ide teda o šetrenie miesta na disku) **a aby sa dali opačným postupom opäť odkódovať**. Existuje mnoho rôznych komprimačných techník.

Komprimácia sa vykonáva buď **automaticky** (uložením súboru v komprimovanom formáte JPG, JPEG, GIF, MPEG, MP3) alebo pomocou **špeciálneho komprimačného programu**. K najznámejším komprimačným programom patria: WinZIP, WinRAR, 7zip, PowerArchiver, FilZip, Windows Commander. Komprimačný program môže vytvoriť niekoľko druhov výstupných súborov:

- bežný archívny súbor - kapacita obmedzená len voľným miestom na cieľovom disku
- delený archív - vznikne rozdelením jedného veľkého archívu na niekoľko častí s danou maximálnou veľkosťou, napríklad pre uchovanie archívu na niekoľkých disketách (vhodné pre kapacitu niekoľko desiatok MB)
- samorozbalovací archívny súbor (*.exe)- SFX (SelF-eXtracting) archív je archív s pripojeným spustiteľným modulom, ktorý po spustení dokáže súbory z archívu rozbaliť. K rozbaleniu SFX archívu nie je teda treba žiadny externý program - sám vykoná dekompresiu do aktuálneho adresára

Vo všeobecnosti (princíp fungovania kompresných algoritmov): **Kompresné algoritmy** obvykle hľadajú v komprimovaných údajoch často sa opakujúce vzory, ktoré sú na výstupe reprezentované jedným opakovaním vzorky a počtom opakovaní. Kompresné algoritmy vytvárajú tzv. **komprimované formáty**.

Napr. *pri kompresii obrázka* sa využíva to, že sa v ňom často opakujú dlhé úseky bodov rovnakej farby. Takýto úsek sa v kóde nahradí iba údajom o farbe a počte rovnakých bodov.

Napr. *pri komprimácii textu*, v ktorom sa 30krát vyskytuje slovo archivácia (toto slovo zaberá $30 \times 10 = 300$ znakov) sa namiesto slova archivácia použije nejaká značka, napr. #, ušetrí sa tým $30 \times 9 = 270$ znakov. Pri komprimácii sa teda každé slovo archivácia nahradí značkou # a do tabuľky kódov sa zaznačí: # = archivácia. Tým sa výrazným spôsobom zmenší veľkosť textového súboru. Komprimačné programy pritom pracujú na ešte nižšej úrovni. Namiesto opakovania znakov sledujú opakovanie bitov (0 alebo 1).

Pomer veľkosti pôvodného a skomprimovaného súboru sa nazýva **kompresný pomer**. Udáva sa prostredníctvom **pomeru** (1:4 (4:1) - skomprimovaný súbor je 4-krát menší ako pôvodný)

alebo v **percentách** (77 % - nový súbor má veľkosť 77 % pôvodného). Komprimovať dáta má zmysel iba vtedy, ak je kompresný pomer menší ako jedna.

PR: 20 MB súbor má po komprimácii 5 MB, teda zmenšil sa na $\frac{1}{4} = 25\%$ pôvodnej veľkosti.

Na základe spôsobu komprimovania údajov možno kompresiu rozdeliť na:

Bezstratová komprimácia – nedochádza k strate údajov, t.j. umožňuje rekonštruovať skomprimovaný súbor do pôvodnej podoby (dekomprimáciou získame úplne identické dáta) – ide o štandard, ktorý zrejme očakávame pri komprimovaní dokumentov, databáz a pod. Kompresia sa deje na základe vynechania redundantných (nadpočetných) informácií.

Stratová komprimácia prostredníctvom komprimácie niektoré údaje z pôvodného súboru vypustí a pri spätnej rekonštrukcii (dekomprimácii) ich už nedokáže obnoviť. Typické použitie je pri grafických súboroch (jpg, jpeg), videu (mpeg), zvuku (mp3). Využíva skutočnosť, že pri porovnaní originálneho a skomprimovaného obrazu či zvuku človek nevníma rozdiely medzi jemnými farebnými prechodmi a používa sa na miestach, kde je možné zánik niektorých informácií tolerovať a ich strata je vyvážená významným zmenšením súboru. Napriek niekoľkonásobnej redukcii dát je výsledok komprimácie takmer nerozoznateľný od orifinálu.

Aké súbory sa oplatí komprimovať?

- Výhodou archivačných formátov je, že dokážu **do jedného balíka** vložiť **množstvo súborov** čo je výhoda, ktorú určite oceníme napr. pri mailovaní.
- Ak však archivujeme súbory len **za účelom zníženia ich veľkosti** je dobré si uvedomiť, že komprimovať sa oplatí len súbory, na ktorých ešte **nie je aplikovaná žiadna kompresia**.

Aj preto majú napr. obrázky vo formáte TIFF alebo BMP po pridaní do archívu veľmi malú veľkosť, zatiaľ čo obrázky v JPEG nezmenšíme takmer vôbec. To isté platí aj o hudobných súboroch. Veľmi efektívne je komprimovať textové dokumenty, databázy, súbory s poštou, objemné savy z hier, jednoducho súbory kde je predpoklad výskytu redundancie.

Základné operácie komprimačných programov sú:

- komprimácia (pakovanie), dekomprimácia (rozpakovanie)
- vytvorenie archívu – vytvorenie jedného súboru z viacerých súborov
- rozbaľovanie archívu (extrakcia)
- pridanie do archívu, odobratie z archívu, prehliadanie obsahu archívu, používanie hesiel, možnosť ovplyvniť mieru komprimácie, ...
- so skomprimovaným súborom **nie je možné vykonávať operácie** ako čítanie, zapisovanie. Na to, aby bolo možné skomprimovaný súbor meniť, je potrebné ho **dekomprimovať**, čiže **rozpakovať**.

Dekomprimácia je opačný proces ako komprimácia. Napr. namiesto značky # sa v komprimovanom texte vloží slovo archivácia.

Úlohy:

1. Akú veľkosť bude mať 15 MB súbor po kompresii s pomerom 2 : 5?
2. Po skomprimovaní mal súbor veľkosť 9 kB. Aká bola jeho pôvodná veľkosť, ak kompresný pomer je 1 : 3?
3. Aký je kompresný pomer v kompresii, ak sa nám 3,6 GB súbor podarilo zmenšiť na 2,4 GB?
4. Vytvorte rozsiahly textový dokument (nakopírujte tento učebný materiál 10 krát za sebou do jedného súboru), zapíšte si jeho veľkosť. Potom ho skomprimujte. Zapíšte si veľkosť po komprimácii. Vypočítajte kompresný pomer použitej kompresie. Experimentujte s rôznymi kompresnými algoritmami.