

# Archivační a komprimační programy

## Kompres

### Nejběžnější archivační a komprimační programy

#### Stručný popis algoritmů bezztrátové komprese

I když paměťová média mají stále větší kapacitu a neustále klesá cena uložení jednotky informace, potřeba úsporného ukládání dat je stále aktuální. Jde zejména o hospodárnější distribuci programů a dat a o lepší využití telekomunikačních kanálů. Navíc moderní techniky ztrátové komprese umožňují vznik zcela nových aplikací jako je digitální záznam videa či přenos telefonních a jiných zvukových zpráv.

**Algoritmy na redukci objemu dat** se dají rozdělit na dvě velké skupiny:

- **Bezztrátová komprese**, kdy komprimovaná data po dekomprimaci jsou naprosto totožná s daty před komprimací, nedochází tedy ke ztrátě informací.
- **Ztrátová komprese** se uplatňuje u obrazových a zvukových dat. Výsledek po dekomprimaci komprimovaných dat se sice nerovná původním datům, je jim však velice podobný a s ohledem na nedokonalost lidských smyslů je rozdíl nevýznamný. Cílem je ukládat data úsporně a dosáhnout toho, aby rozdíl mezi uloženými a původními daty byl téměř nepostřehnutelný.

Pro úplnost je třeba dodat, že pokud máme na mysli redukci objemu dat, nesmíme opomenout i redukci na základě textu v přirozeném jazyce a výběru klíčových částí textu. Tento problém je jednou z hlavních oblastí výzkumu v oblasti umělé inteligence, s archivačními a komprimačními programy nemá však nic společného.

### **Bezztrátová komprese**

Bezztrátová komprese je použitelná **pro všechny druhy dat** (texty, obrázky, zvuky apod.) bez rozdílu. Cenou univerzální použitelnosti je však **omezená účinnost**, která se vyjadřuje poměrem velikosti originální a redukované verze dat, jež v průměru dosahuje zhruba 2 : 1 (výsledek pochopitelně závisí na konkrétních datech).

Bezztrátová komprese nachází uplatnění v řadě oblastí. V dalším textu jsou tyto oblasti popsány.

### **Zálohování a archivace**

Zálohování a archivace se provádí na pásku či disk pomocí **zálohovacích a archivačních programů**, které mají díky kompresi šetřit prostor na médiu a čas potřebný pro uložení dat, neboť zálohovací média jsou obvykle pomalá. Kompresce se provádí pomocí programů užívajících univerzální kompresní algoritmy a též hardwarově na řadiči příslušného zařízení (streameru).

Smyslem zálohování a archivace je především vytvářet záložní média a záložní data pro případ obnovy omylem smazaných dat či havárie. Zálohovací programy poskytují takové služby jako je archivace jen těch souborů, které od poslední archivace doznaly změny, automatická archivace po uplynutí určité doby, odložené zálohování dat na síťových serverech v době relativního klidu (o půlnoci) apod. V profesionálním nasazení počítačů se užívají speciální zálohovací programy.

Pro zálohování se používá rozšířených **utilit** (PkZip, Arj, Rar a dalších), které jsou popsány v dalším textu. Jejich užití je jednoduché, jsou dostupné (řada z nich jsou sharewary) a

některé jsou pro nekomerční účely použitelné i zdarma. Umožňují vytvářet archivy, do kterých lze přidávat další soubory, starší kopie souborů nahrazovat novějšími, rušit archivní kopie souborů, pro potřeby archivace na disketách lze vytvářet i více diskové archivy apod.

Často se užívají jako dostupná (a méně pohodlná) náhražka zálohovacích programů. Protože však často umožňují i vytváření **samorozbalovacích archivů**, což jsou archivy připojené ke krátkému EXE programu, zajišťujícímu automatickou dekompresi po spuštění uživatelem, užívají se též pro distribuci programů a dat.

Pro distribuci dat se však užívají speciální programy zaměřené výhradně **na distribuci dat**. Užívají se především v instalačních programech. Obsahují různé „triky“, které za cenu silně zpomalené komprese (to však vzhledem k účelu vůbec nevedí) umožňují výrazně zvýšit kompresní poměr.

### Přenos dat

Komprese dat je obvykle přímo vestavěná do používaných komunikačních prostředků. Vzhledem ke stále omezeným kapacitám telekomunikačních linek je komprese při přenosu dat velmi důležitá. Komprese je prováděna jednak technickými prostředky a jednak programovou nadstavbou. Modemy mají kompresní metodu přímo zahrnutou v protokolech. Pokud se však přenášejí data komprimovaná již jinou metodou, k žádné další kompresi již nedojde.

### Komprese obrazových dat

Obrazová data jsou přímo vhodná pro kompresi, právě proto se zde datová redukce užívá nejčastěji. Bitmapové obrázky zejména většího rozměru mají nároky na délku souboru téměř neúnosné a přitom datová redukce je zde obvykle velice účinná.

Užívá se různých metod, které vycházejí z různých nároků obrazů. Zcela odlišnou charakteristiku mají např. černobílý fax, ruční kresba nebo digitalizovaná fotografie. Užívají se jak metody bezztrátové, tak i ztrátové.

Pro barevné fotografie ve vysokém rozlišení se často užívají ztrátové metody. Pro obrázky černobílé či s malou paletou barev (do 256) se hodí metody bezztrátové. Pro obrázky typu ruční černobílá kresba nebo fax se hodí jednoduché komprimační metody, protože tyto obrázky obsahují velké bílé plochy.

V komprimaci barevných obrázků vede algoritmus LZW, který je využíván v řadě grafických formátů (zejm. TIFF a GIF), jeho nevýhodou je však jeho patentová ochrana. Zejména z těchto důvodů byl vyvinut nový grafický formát PNG (Portable Network Graphics).

### Ztrátová komprese

Ztrátová komprese se užívá pro záznam fotografií ve vysokém rozlišení, přenos zvuku či videa. Vzhledem k vysokému rozlišení by byl výsledný objem informací tak rozsáhlý, že by bylo prakticky nemožné tyto informace ukládat a zpracovávat. Vyvíjejí se proto techniky ztrátové komprese, které respektují vlastnosti přenášených dat a způsoby jejich vnímání lidskými smysly. Vývoj komprese zejména v oblasti audio a videozáznamů je překotný, což souvisí se snahou přenášet tyto záznamy přes globální komunikační sítě. Firmy výsledky svého snažení většinou nepublikují a ochraňují si je patentově či licenčně .

Nejjednodušší typy ztrátové komprese se užívají u fotografií. Využívá se zde nedokonalostí lidského oka podobně jako u řešení barevné televize. Metoda fraktálové komprese „hledá“ v obrázku útvary podobné fraktálovým obrazcům, které lze popsat několika málo para-

metry. Využívá toho, že řada objektů reálného světa vykazuje soběpodobnost - má „fraktálových charakter“.

Významnou vlastností ztrátové komprese je možnost široce parametrizovat redukci a získávat tak různé kvalitní výsledky. Lze tedy si předem stanovit procento ztráty informace, výsledek redukce si prohlédnout, posoudit a vybrat si pak redukci, která přináší nejlepší výsledky z hlediska kompresního poměru a stupně zkrácení obrázku. Těchto způsobů užívá např. **obrazový formát pro ukládání fotografií JPEG**.

### Multimédia

U **zvukových dat** se při redukci vychází z frekvenční analýzy a vhodným způsobem se vybírají jen ty frekvence, které jsou pro lidské ucho nejdůležitější. Vždy je důležitý cíl - je to buď maximální redukce záznamu bez nároků na vyšší kvalitu, anebo zachování co nejvyšší kvality při stanovené maximální horní hranice kapacity přenosového kanálu.

Nejnáročnější je komprese pohyblivého obrazu se zvukovým doprovodem (video sekven- ce). Vychází se zde z toho, že sousední obrázky v rámci jednoho záběru se liší jen na ome- zené ploše, vybere se proto „zajímavá oblast“ obrázku a aplikuje se ztrátová komprese. Tuto technologii pochopitelně nelze aplikovat na hranici dvou záběrů nebo při prudké změně scé- ny. Protože však lidské oko je méně citlivé než ucho, krátké pozdržení předchozího obrazu neruší tak jako výpadek zvuku. To ovšem klade vysoké nároky na udržování plynulého zvu- kového doprovodu při přehrávání digitálního videa.

Některé metody ztrátové komprese se uplatňují pro kompresi i dekompresi v reálném čase, jiné potřebují delší dobu komprese, která se musí provádět odděleně. Platí to pře- devším pro **kompresi videa**, která je výpočetně velice náročná. Příkladem takové komprese je **MPEG**.

## Nejběžnější archivační a komprimační programy

### ARJ

Ve svém názvu skrývá iniciály svého původního autora Roberta Junga.

ARJ je značně rozšířen a oblíben. K jeho provozování stačí jediný soubor ARJ.EXE. Kompletní dodávku však tvoří více souborů (kompletní dokumentace, zjednodušené verze ARJ aj.). ARJ lze používat na počítači jen na příkazové řádce. Některé souborové manažery (Norton Commander či Windows Commander) umějí s ARJ archivy zacházet přímo. ARJ umožňuje mj. třídění souborů v archivech, vytvářet archivy na disketách, tvorbu samoroz- balovacích archivů, zabezpečit archiv heslem, ověřovat úspěšnost archivace aj. Existuje uti- lita ArjView.exe, která usnadňuje prohlížení archivu a jeho dekomprimaci.

Ve srovnání s nejnovějšími programy je ARJ komprimačním poměrem průměrný, v rych- losti komprimace patří mezi lepší a rychlostí dekomprimace se řadí mezi nejlepší.

### PkZip

Archivy ZIP jsou ještě rozšířenější než ARJ. Pracuje se zde se dvěma programy: PkZIP pro komprimaci a PkUNZIP pro dekomprimaci dat.

PkZIP je starší než ARJ. Tento formát je podporován mnoha utilitami pro téměř všechny operační systémy: DOS, Windows, OS/2, Unix, Amiga a Atari.

Uživatelským komfortem je srovnatelným s ARJ. I tento program se spouští na příkazové řádce. Chudší výbavu než ARJ má pro vytváření disketových souborů. Samorozbalovací ar- chivy lze vytvářet jen pomocí externí utility. Výkonností je přibližně srovnatelný s ARJ.

## WinZip

Je specializovaný program pro podporu formátu ZIP v prostředí Windows. Zaujme uživatelským komfortem, hezkým grafickým prostředím a nápaditými ikonami. Po jeho instalaci do Windows přibude v Průzkumníkovi (Explorer) po kliknutí pravého tlačítka myši na soubor či složku položka Přidej do ZIP archivu (Add To Zip). Důležitou volbou je ukládání dlouhých jmen souborů ve formě akceptovatelné i pro jiné prostředí než Windows. Je díky Windows o něco rychlejší a účinnější než PkZip.

## RAR

Tento program je relativně nový, navazuje však na dobré tradice. Všechny jeho verze lze provozovat jak z příkazové řádky, tak i v integrovaném uživatelském rozhraní. Základní verze pro DOS je svým vzhledem inspirována Norton Commaderem. Uživatelsky je program i v prostředí DOS uživatelsky přátelský. Služby jsou srovnatelné s ARJ, velmi snadné je vytváření samorozbalovacích archivů. RAR zavádí nový typ archivu zvaný Solid optimalizovaný pro archiv s velkým počtem malých souborů. RAR je velmi dobrý pro komprimaci multimediálních souborů, pro které má speciální podporu. Celkově je velmi dobrým archivátorem. Instalace do Windows 9x však není tak dobře propracovaná jako u WinZip. Rychlost v prostředí 32 bitů není významně vyšší než u verze pro DOS. Existuje i jeho verze WinRAR.

## Stručný popis algoritmů bezztrátové komprese

### Proudové kódování

Je metoda založená na zjišťování opakujících se znaků. Je to algoritmus jednoduchý, rychlý, avšak s nikterak výraznými výsledky. Je vhodné pro černobílé obrázky s velkými bílými plochami a obecně pro soubory se rozsáhlejšími bloky opakujících se hodnot. Kromě starších grafických formátů se užívá jako prvotní postup před uplatněním jiného, efektivnějšího algoritmu. Značně zjednodušená demonstrace: opakující se znaky v posloupnosti se nahradí příznakem opakování, počtem výskytů v opakování a opakujícím se znakem. Např. AAAABCDEEEF se zakóduje jako \*4ABCD\*3EF, kde hvězdička představuje příznak opakování. Výhodou proudového kódování je velmi jednoduché dekódování.

### Kódování opakovaných řetězců

Principem je, že se vyhledávají řetězce (posloupnosti znaků) v předchozích datech, pokud byl již řetězec užít, místo jeho opakování se uvedené pouze kód tohoto řetězce. Příkladem jsou metody zvané **Lempel-Ziv** algoritmy (označované LZ...) nazvané podle svých tvůrců (Lempela a Ziva). Např. LZ77 si lze představit tak, že se postupně prochází vstupní text a narazí-li se na opakující se řetězec, pak se místo jeho opakování vloží příznak opakovaného řetězce, pozice a délka řetězce, který se opakuje. Pro představu: řetězec ABCBCACB se kóduje jako ABC\*22A\*32. Hvězdička je opět příznak opakování, první číslice pozice a druhá délka opakujícího se řetězce.

### Huffmanovo kódování

Tato metoda využívá různé četnosti znaků ve zdrojovém textu. Zatímco běžné kódování znaků v počítači přiřazuje každému znaku stejný počet bitů (zpravidla 8), zde se nejčtenější znaky uloží s minimálním počtem bitů a méně čtené pak se stále větším. Celkově se tak průměrná délka daná počtem bitů zobrazujících jeden znak minimalizuje. Např. v posloupnosti ABCBADBEAB je nejčtenější B (4 výskyty), dále A se třemi výskyty a konečně C, D, E po jednom výskytu. V tomto případě by pro B stačil jeden bit, pro A dva a pro C tři a pro D, E

po čtyřech. Pro celkovou délku by pak stačilo 21 bitů a nikoli 10 znaků \* 8 bitů, tj. 80 bitů v kódu ASCII.

Podstatou kódování je, že text se poprvé podrobí statistickému zjišťování počtu znaků a vytvoří se posloupnost znaků (dlouhá např. 256 bajtů pro 256 znaků) uspořádaných podle výskytu znaků v kódovaném souboru. Tato posloupnost se pak pro účely dekódování připojí k výslednému zakódovanému souboru. Při druhém průchodu souborem se pak provádí kódování. Z uvedeného je zřejmé, že velmi krátké soubory se nekódují, protože výsledný soubor by byl delší než původní.

Huffmanovo kódování se obvykle užívá v kombinaci s jinými metodami a to jak pro bezztrátovou kompresi, ale i pro kompresi ztrátovou. Např. PkZip i Arj užívají kombinaci opakujících se řetězců (LZ77) a Huffmanova kódování.

### **Literatura:**

Téma týdne - Computerworld 11/97