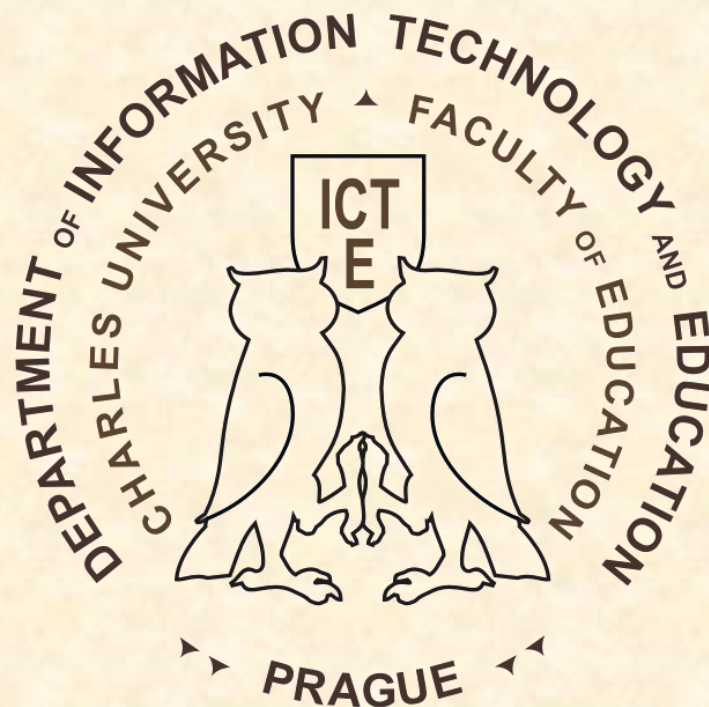




Univerzita Karlova, Pedagogická fakulta
Katedra informačních technologií a technické výchovy

Kódování, komprese



Kódování

- Je způsob reprezentace dat, informací a hodnot při jejich uložení v paměti, v souboru, při přenosu po síti atd.
- Příkladem může být ASCII kód, který písmena a jiné znaky reprezentuje jako čísla uložená na jednom bytu (8b).

Kódování

- **Kód** – nahrazení jedné abecedy (*abeceda zdroje*) jinou abecedou (*kódová/přenosová abeceda*). Např. Morseova abeceda (1840).
- Usnadnění technické stránky záznamu, zpracování nebo přenosu zprávy.
- Kód může být graficky zapsán ve znacích stejné abecedy jako původní zpráva (např. 0,1).

Kódování

- Kódování by nemělo „komplikovat život“ nebo „zatěžovat zařízení“; úspornost kódu.
- *Kódová slova nejsou stejně dlouhá*
 - Znaky/slova abecedy zdroje, které se vyskytují častěji kódovány kratšími kódovými slovy; kód s proměnlivou délkou slova.
- *Kódová slova stejně dlouhá*
 - Délka n (bitů); neměla by být delší, než je nutné (např. ASCII), blokový kód.

Kódování

- *Prefixový kód* - žádný symbol jeho kódové abecedy není předponou (prefixem, začátkem) jiného (dalšího) symbolu abecedy (např. Huffmanovy kódy). Dekódování zleva doprava.
- *Neprefixový kód* – použití oddělovačů (např. Morseova abeceda).

Komprese (komprimace)

- Zpracování počítačových dat s cílem *zmenšit jejich objem* při současném zachování informací v datech obsažených.
- *Zmenšit datový tok* při jejich přenosu nebo zmenšit potřebu zdrojů při ukládání informací.
- Kódováním, které je dané zvoleným kompresním algoritmem, se ze souboru odstraňují *redundantní* (nadbytečné) informace (zvyšuje se entropie dat), *irelevantní* data.
- Proces aplikace kompresního algoritmu.

Komprese

- **Ztrátová komprese** – při kompresi jsou některé informace nenávratně ztraceny a nelze je zpět zrekonstruovat (nevýhoda určitého zkreslení je vyvážena významným zmenšením souboru).
 - Používá se pro kompresi zvuku a obrazu/videoa.
- **Bez(e)ztrátová komprese (přesná, vratná)** – obvykle není tak účinná jako ztrátová komprese; komprimovaný soubor lze opačným postupem zrekonstruovat do původní podoby.

Kvalita komprese

- Rychlost komprese.
- Symetrie/asymetrie kompresního algoritmu.
 - **Symetrické algoritmy** – stejný čas potřebný pro kompresi i dekompresi.
 - **Asymetrické algoritmy** – čas potřebný pro kompresi a dekompresi se liší.
- Kompresní poměr.

Kompresní poměr

- Je poměr délky zkomprimovaných dat (výstupní) vzhledem k délce původních dat (vstupní).
- Kompresní poměr se často vyjadřuje v procentech.
- Kompresní poměr je ovlivněn volbou kompresního algoritmu i typem komprimovaných dat.
- Kolik je při kompresi v průměru zapotřebí bitů pro uložení jednoho bytu výchozího souboru (**bpb – bits per byte**).
- Úspora místa je vyjádřena jako „*1 - velikost komprimovaných dat / velikost nekomprimovaných dat*“

Př. Soubor 10MB před kompresí, po komprimaci 2MB => $10/2=5$ (5:1)
=> $1-1/5=0,8$ => 80% úspora

Kompresní metody

- **Statistické metody komprese** - jsou založeny na pravděpodobnosti, s jakou se vyskytují jednotlivé znaky v textu (Huffmanovo kódování, Aritmetické kódování).
- **Slovníkové metody komprese** - principem těchto metod je vyhledání opakujících se částí textu. Do zkomprimovaného textu se uloží jen první výskyt takové části. Všechny další výskyty jsou nahrazeny odkazem na předchozí výskyt (LZW).
 - Jednoduché - kódování opakujících se posloupností znaků (RLE).
- Komprese nepohyblivého obrazu, komprese pohyblivého obrazu, komprese zvuku.

Slovníkové metody

- Jsou založeny na vyhledávání opakujících se částí textu - **fráze**.
- Při kompresi se do výstupního souboru uloží jen první výskyt části textu a další výskyty se nahradí odkazem:
 - První třída metod nahrazuje výskyt opakující se části textu odkazem na místo, kde se v textu tato část předtím vyskytla.
 - Druhá třída si z částí textu, které se předtím již vyskytly, vytváří v paměti slovník a další výskyty nahrazuje odkazem na příslušné místo ve slovníku.

Slovníkové metody - RLE

- Run Length Encoding (kódování délkou běhu).
- Kódování délkami sekvencí.
- Proudová – na vstupu jednotlivé znaky/symboly.
- Jednoprůchodová.
- Symetrický algoritmus (stejná časová složitost).
- Kódování opakujících se posloupností znaků.
- Za sebou více následujících stejných hodnot kóduje do dvojice hodnota a počet.
- Sekvence kratší než 3 znaky včetně, ukládá na výstup bez použití indikátoru komprese; escape character (únikový znak).
- Nevhodné – text, vhodné – obrázky (PCX, BMP).

Slovníkové metody – LZ77

- Lempel, Ziv – 1977.
- Nahrazuje výskyt opakující se části textu odkazem na místo, kde se v textu tato část předtím vyskytla.
- Postupem času celá třída kompresních metod LZ77.
- Metoda s adaptivním modelem.

Slovníkové metody – LZ77

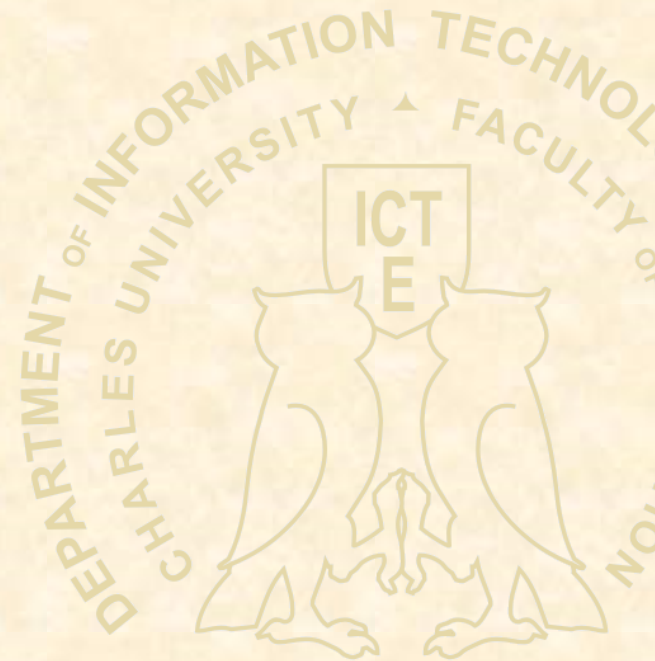
- Jednoprůchodové - nevyžaduje předběžnou analýzu souboru, asymetrická (rychlejší dek.).
- Koncept výhledového okénka (sliding window=posuvné/klouzavé okno).
- Odvozen LZSS (Lempel–Ziv–Storer–Szymanski)
 - LZSS + Huffman. kód. používá řada kompresních programů (ZIP, ARJ, RAR).
- Adaptivní model.

Slovníkové metody – LZ78

- Lempel, Ziv – 1978
- Z částí textu, které se předtím již vyskytly, vytváří v paměti slovník a další výskyty nahrazuje odkazem na příslušné místo ve slovníku.
- Postupem času celá třída kompresních metod LZ78.
- Metoda s adaptivním modelem.
- Obrázky, texty.

Slovníkové metody – LZW

- Lempel-Ziv-Welch (1984)
- Odvozená od LZ78
- Jednoprůchodová
- GIF, TIFF, ARJ, ZIP



Kompresní metody statistické

- *Huffmanovo kódování, aritmetické kódování* (např. Q-kódování)
 - Kompresní poměr dosahovaný při použití Huffmanova i aritmetického kódování je nízký.
 - Jako hlavní kompresní metoda se nepoužívá.
 - Efektivní prostředek k uložení číselných nebo znakových hodnot, které jsou výstupem jiných, dokonalejších kompresních metod.

Vstupní text -> základní kompres. metoda -> čísla/znaky
-> Huff./Aritm. kódování -> výstup

Huffmanovo kódování (1952)

- Metoda založená na stanovení četnosti výskytů jednotlivých znaků v kódovaném souboru a kódování znaků s největší četností slovem s nejkratší délkou.
- Vytvoření kódu s minimální redundancí.
- Huffmanova konstrukce minimálního kódu je vždy optimální, je jednoznačná, ale není jedinečná.

Huffmanovo kódování

- Model dat statický/dynamický.
- Statistický algoritmus (četnost znaků).
- Dvouprůchodový (analýza, poté komprese).
- Kódovací strom – binární strom (prefixový kód).
- Asymetrický vstup/výstup (dekom. rychlejší než kom.).
- Kódové symboly mají proměnnou délku.
- Běžný text, obrázky.

Model dat

- Údaje o pravidelnostech a redundancích v komprimovaném textu nazýváme jeho modelem.
- Model je východiskem pro příslušnou kompresní metodu.
- U metod statistického kódování jsou modelem pravděpodobnosti (nebo frekvence) výskytu jednotlivých znaků v textu.
- Huffmanův kód podle modelu stanoví délku kódování každého znaku.

Model dat - statický

- Pevně stanoven při implementaci kompresní metody.
- Jednotný model používaný pro kompresi všech vstupních textů; kodér/dekodér používají stejný model dat.
- Účinnost komprese je ale závislá na tom, do jaké míry komprimovaný text odpovídá modelu - čím více se text od modelu liší, tím více se snižuje účinnost komprese (nižší účinnost kódování, neboť se ve větší míře kóduje znak s delším kódem).
- Doplnující způsob kódování, rychlejší.
- Pro velké množství dat neměnný – texty ve velké knihovně.

Model dat - semiadaptivní

- Přizpůsobení modelu vstupním datům.
- Způsob, kdy kompresní program si nejprve před kompresí ze vstupního textu vytvoří model (Huffmanův strom) a pak s tímto modelem text zkomprimuje.
- Model musí přidat ke zkomprimovanému textu, neboť dekompresní program v tomto případě není schopen model sám vytvořit.
- Huffmanovo kódování - kompresní program nejprve zjistí frekvence výskytu znaků v textu a z nich vytvoří strom s Huffmanovým kódem.
- Doplnující způsob kódování, rychlejší.

Model dat – adaptivní (dynamický)

- Způsob, kdy kompresní program si model vytváří až v průběhu komprese podle vstupního textu (během čtení vstupních dat); model se vyvíjí.
- Model je zapotřebí nejen pro kompresi, stejný model je rovněž potřebný při dekompresi.
- Asymetrické kódování.
- Největší účinnost komprese - dominantní kompresní metody jsou založeny na adaptivním modelu.