Vyhodnotenie práce na ročníkovom projekte za zimný semester

Autor: Ondrej Škorňák Dátum: 25. 1. 2024

Úvod:

V tomto ročníkovom projekte sa snažím zameriavať na štúdium a implementáciu tunelovacej metódy pre blokový kompresor Bzip2. Cieľom je vytvoriť plne funkčný kompresný algoritmus, ktorý by mohol zlepšiť kompresný pomer a efektivitu kompresie.

Ciele projektu:

Môj projekt má dva hlavné veľké ciele:

- Štúdium BWT (Burrows-Wheeler transform) a jeho využitia v kompresii.
- Implementácia tunneling metódy pre blokový kompresor Bzip2 s cieľom vytvoriť plne funkčný kompresný algoritmus.

Počas tohto semestra som sa snažil plne zamerať na prvý z týchto dvoch cieľov. Na dosiahnutie tohto cieľu som si stanovil tri podciele pre zimný semester:

- Naštudovať základné algoritmy a štruktúry používané v kompresii dát.
- Porozumieť metóde tunelovania a jej aplikácii v B2zip.
- Analyzovať možnosti zlepšenia miery kompresie pomocou tunelovania.

Metodika:

V prvom rade som sa sústredil na štúdium základných algoritmov a štruktúr spojených s kompresiou. Keďže som sa nikdy predtým nestretol s podobnou tematikou, tak som si chcel vytvoriť dobré základy aj pre budúce nadväzné štúdium a byť si istý, že budem schopný pochopiť o čom je technika tunelovania a algoritmus B2zip.

Preto som si naštudoval algoritmy ako sú suffix arrays, Burrows-Wheeler transform (BWT), source encoding, Huffman encoding, arithmetic encoding, run-length encoding, move-to-front transform, wavelet trees, tries a wheelerove grafy a ďalšie.

Ako jeden z hlavných zdrojov môjho štúdia slúžila Baierova práca. Táto práca obsahuje všetky relevantné informácie ako aj kvalitné zdroje k hlbšiemu štúdiu algoritmov.

Baier, U. 2020. BWT TUNNELING. Univerzita Ulm.
https://oparu.uni-ulm.de/xmlui/bitstream/handle/123456789/35280/baier_thesis.pdf?sequence=3

Dosiahnuté výsledky:

- Naštudoval som základné algoritmy a štruktúry potrebné pre pochopenie a implementáciu kompresie.
- Pochopil som metódu tunnelingu a jej komplexnosť, vrátane NP-ťažkosti plánovania tunnelingu.
- Identifikoval som rôzne stratégie implementácie tunnelingu, vrátane greedy stratégie navrhnuté Baierom.

- Analyzoval som potenciálne zlepšenie miery kompresie v B2zip vďaka tunnelingu, najmä pri repetitívnych dátach.

Bzip2 je voľne dostupný a patentom neobmedzený algoritmus pre bezztrátovú kompresiu dát. Bol vytvorený Julianom Sewardom a prvýkrát vydaný v roku 1996. Bzip2 je založený na Burrows-Wheelerovej transformácii (BWT), ktorá preorganizuje textové reťazce do takých, ktoré sú lepšie komprimovateľné pomocou techník ako je run-length encoding (RLE) a Huffmanovo kódovanie. Bzip2 je obzvlášť efektívny pri kompresii veľkých súborov a súborov s opakujúcimi sa sekvenciami.

Tunelovanie je technika, ktorá sa používa v kontexte BWT na zlepšenie kompresie. Táto technika spočíva v identifikácii a spojení "tunelov" v Wheelerovom grafe, ktorý je reprezentáciou BWT. Tunel v tomto kontexte je sekvencia znakov, ktorá sa opakuje v pôvodných dátach. Spojením týchto tunelov sa znižuje redundancia v dátach a zlepšuje sa kompresný pomer.

Tunelovanie je dôležité pre zlepšenie kompresie, pretože umožňuje efektívnejšie využitie existujúcich kompresných techník. V prípade Bzip2 môže tunelovanie výrazne zlepšiť kompresný pomer, najmä pri kompresii veľkých súborov alebo súborov s vysokou mierou redundancie. Avšak, implementácia tunelovania môže byť náročná a môže zvýšiť časovú a pamäťovú náročnosť kompresie.

Výzvy a problémy:

Najväčšou výzvou zatiaľ bolo samotné pochopenie metódy tunelovania. A myslím si že najväčšou výzvou v letnom semestri bude aj jeho efektívna implementácia.

Navrhované riešenia:

Myslím si s mojimi doposiaľ nadobudnutými skúsenosťami, že by implementácia tunelovania mohla vyzerať nasledovne.
Začlenenie tunelovania do schémy Bzip2 by zahŕňalo úpravu fázy BWT.
Modifikovaná schéma s tunelovaním by mohla vyzerať:

1. Run-length encoding (RLE) na počiatočných údajoch
2. Burrows-Wheeler transform (BWT) s tunelovaním:
    a. Vykonajte štandardnú BWT na triedenie prípon a vytvorenie reťazca BWT.
    b. Reprezentujte BWT ako Wheelerov graf.
    c. Aplikujte tunelovanie na Wheelerov graf spájaním paralelných rovnako označených ciest.
    d. Vytvorte tunelovaný reťazec BWT z tunelovaného Wheelerovho grafu.
3. Move-to-front transform (MTF)
4. Run-length encoding (RLE)
5. Huffman coding a ďalšie fázy v b2zipe

Diskusia:

Implementácia tunelovania by mohla priniesť lepšiu mieru kompresie, ale zároveň by mohla zvýšiť zložitosť implementácie, zvýšiť časovú náročnosť a pamäťové nároky, a vyžadovať úpravy dekompresných algoritmov.

Záver:

Téma kompresie dát a špecificky implementácie tunelovacej metódy v B2zip ma veľmi zaujala. Je to moje prvé stretnutie s kompresnými algoritmami, čo mi prináša nové výzvy a príležitosti na zlepšenie mojich technických zručností.

Čo ma na tomto projekte najviac zaujalo, je potenciálny reálny dopad (aj keď malý), ktorý môže mať. V dnešnom svete, kde sa generuje obrovské množstvo dát, je neustále zlepšovanie a vývoj kompresných algoritmov kľúčové. Rýchlejšie a efektívnejšie kompresné metódy môžu mať významný vplyv na rôzne oblasti, od ukladania dát až po ich prenos.

Doposiaľ v rámci tohto projektu som sa naučil mnoho nových konceptov a techník, ktoré mi otvorili nové perspektívy na pochopenie a riešenie problémov v oblasti kompresie dát. Tieto zručnosti a poznatky mi pomôžu v mojej ďalšej kariére a budú ma motivovať k ďalšiemu štúdiu a výskumu v tejto oblasti.

V blízkej budúcnosti plánujem pokračovať v práci na tomto projekte, kde sa budem zaoberať programovaním a experimentovaním so zdrojovým kódom B2zipu. Som presvedčený, že tieto aktivity mi umožnia lepšie pochopiť a optimalizovať tunelovaciu metódu a jej implementáciu v B2zip. Taktiež sa teším na možnosť testovať a hodnotiť výsledky mojej práce, čo mi poskytne cenné spätné väzby a pomôže mi zlepšiť moje riešenie. Teším sa na ďalšie kroky v tomto projekte a som odhodlaný pokračovať v práci na dosiahnutí hlavných ako aj vedľajších cieľov projektu, ktoré som si stanovil na začiatku.