

Posudek oponenta diplomové práce

**Stanislav Kovalčín:**

## *Slovníkové metody jako druhá fáze BWT*

Autor ve své práci testuje, jaký vliv mají vybrané fáze komprese XBW vycházející z Burrows-Wheelerovy transformace. Věnuje se zejména různým variantám MTF, RLE a slovníkové komprese – tedy fázím bezprostředně následujícím za BWT.

Připravil různé varianty těchto kroků (včetně jejich vypuštění) a měřil, jak se která kombinace chová. Překvapilo mě, že nebyla provedena rozvaha, které kombinace principiálně nabízejí něco zajímavého a kde by byl výrazný zisk překvapivý – např. kombinace RLE a slovníkové komprese nabízí domněnku, že použití RLE nepovede k velkým ziskům, neboť úpravy proudu dat metodou RLE mohou dokonce zhoršit chování slovníkových metod, což se provedenými pokusy potvrdilo. Uvedenými úvahami na začátku by bylo pravděpodobně možné vytipovat nadějnější kombinace metod.

Autor představuje implementaci RLE pomocí splay-stromu místo klasického spojového seznamu. Důvodem k této úpravě je fakt, že při velké abecedě se prochází příliš mnoho uzlů. Bohužel jsem v práci nenašel zmínku, jak moc nová datová struktura přispěla ke zrychlení komprese či alespoň dané fáze a za jakých protorových nákladů. Mám důvod se domnívat, že tak, jak byl splay-strom navržen, se brzy stává lineárním spojovým seznamem nebo je mu alepoň blízký.

Další připomínku bych měl k zavedení uspořádání, kdy autor požaduje, aby

1.  $x < y \vee y < x$
2.  $x < y \vee y < z \Rightarrow x < z$ ,

což je poněkud nešťastné. (Tato ‘definice’ se v posuzované práci vyskytuje dokonce  $2 \times$  – na stranách 12 a 25.)

Při vysvětlování funkce BWT (str. 13) se do další fáze posílá jak transformovaný řetězec obsahující symbol konce, tak i číslo řádku, kde se původní varianta řetězce nachází. Proč? K rekonstrukci přeci stačí jen jedna z variant určující správný řádek. Na straně 20 se praví, že pro symbol  $x_i$  se vypíše  $F(x)$  místo pravděpodobně zamýšleného  $F(x_i)$ . O kus níže na téže stránce se uvádí, že se vypisují symboly podle pořadí v abecedě – bohužel je to na místě, kde přichází v úvahu abeced více. Obrázky zobrazující uspořádání by jej měly dodržet – to bohužel neplatí na str. 29.

CD obsahuje elektronickou verzi textové části práce (pdf), aplikaci spustitelnou pod Windows a pravděpodobně preložitelnou i pod Linuxem a testovací data. Bohužel se mi nepovedlo najít návod, jak s aplikací zacházet (uživatelskou dokumentaci). Po spuštění se otevřelo okno s aplikací, objevil se prompt a – co teď? Zdá se, že práce – i když v mnoha ohledech náročná a přínosná – vykazuje ještě některé znaky nedokončenosti.

I když je odevzdaná práce na hranici uznatelnosti za práci diplomovou, doporučuji ji k obhajobě.

Praha, 12. září 2007

RNDr. Michal Zemlička, Ph.D.