Oponentský posudek diplomové práce
Tomáše Urbana: Semi-adaptivní slovníkové kompresní metody

Práce je věnována bezztrátové kompresi dat. Jejím deklarovaným cílem bylo navrhnout dvouprůchodovou slovníkovou metodu, která na základě analýzy vstupu vybuduje slovník frází, a ten v dalším průchodu využije ke kompresi. Popis navrženého řešení spolu s výsledky experimentálního vyhodnocení jsou obsahem textu práce. Protože autor pojal celý výzkum jako rozšíření softwarového projektu XBW, elektronickou přílohu práce tvoří kromě zdrojového kódu autora též dokumentace tohoto projektu.

Text práce je napsán stručně a poměrně srozumitelně, protestoval bych pouze vůči systematickému skloňování anglických termínů (práce s trii, abeceda unicodu). Poněkud zmatené je označení v záhlaví tabulek: Jednotky nejsou až na jedinou výjimku uváděny, zkratka „P. slovník" (tab. 4.1 a další) není vysvětlena, záhlaví ostatních sloupců je sice v komentováno v textu, proč ale autor používá střídavě zkratky pro termíny v českém (K.p – kompresní poměr) a anglickém (l výstupu – zkratka pro length?) jazyce? Podobný problém nastává u pseudokódu Algoritmu 4 (str. 24), kde narazíme na identifikátory „délka_zpracovaného_úseku" a „l_výhledu".

Úvodní část textu podává stručný přehled obsahu práce, následovaný popisem programu XBW, který vznikl na MFF jako softwarový projekt a autor na něj navazuje. Bohužel zde chybí jakýkoliv přehled současného stavu zkoumaného problému, tj. alespoň stručná charakteristika existujících slovníkových metod, vysvětlení, že problém konstrukce optimálního slovníku je NP-těžký, a motivace toho, co je v práci zkoumáno.

Zbylá část práce je věnována tvorbě optimálního slovníku. I zde bohužel chybí jakákoliv rešerše současného stavu problému. Autor se zaměřuje výhradně na reprezentaci slovníku pomocí datové struktury trie, popisuje 3 varianty, navržené Lánským a Žemličkou, a u dvou z nich experimentálně testuje prostorovou náročnost. Ve výsledcích (str. 13) nejsou uvedeny žádné časové údaje; autor sice argumentuje, že doba běhu byla pro obě metody stejná, není ale jasné, zda má na mysli čas potřebný pro konstrukci slovníku, pro použití již vytvořeného slovníku k transformaci vstupu, nebo pro jeho úsporné uložení v rámci komprimovaného souboru. Podobná nejistota se vznáší i nad časovými údaji uvedenými v dalších kapitolách.

Jádrem práce je následující kapitola, v níž autor popisuje návrhy pro optimalizaci slovníku. Vylepšení popsaná v odstavci 4.4 jsou téměř triviální, v textu jsem neobjevil žádné důvody, proč se s nimi nepočítalo již od počátku. Mnohem zajímavější jsou úpravy navržené v odstavci 4.5, kde se autor snaží o prořezání stromové datové struktury. Za poznámku stojí, že úprava PT1B odstraní vrcholy s jediným synem, a vede tedy k datové struktuře, známé jako sufixový strom, zde ovšem stále reprezentované jakožto sufixový trie. Je škoda, že autor nezvážil možnost pracovat přímo se sufixovým stromem, který má lineární prostorovou složitost v délce vstupu, což představuje alespoň teoretickou výhodu oproti sufixovému trie se složitostí v nejhorším případě kvadratickou. Všechny navržené modifikace jsou experimentálně testovány a autor se snaží i o interpretaci výsledků.

Závěr práce je věnován experimentálním testům nejúspěšnějších z navržených variant a srovnání se třemi standardními algoritmy bezztrátové komprese. Tabulky 6.1. a 6.2. obsahují i časové údaje, není ale jasné, zdali jsou do časové náročnosti metod PT zahrnuty i fáze tvorby slovníku a transformace vstupu, v předchozím textu (tab. 5.1. a 5.2.) se uvádí zřejmě jen časová náročnost komprese slovníku.

Elektronickou přílohu práce tvoří CD, v jehož obsahu se lze jen velmi obtížně orientovat. Podle informací v textovém souboru POPIS.TXT je zdrojový kód autora uložen v jednom ze 17 podadresářů. Obsah ostatních, o nich se autor nezmiňuje, pochází zřejmě ze softwarového projektu XBW. Tento projekt není uveden v seznamu literatury, takže ani není zcela jasné, kdo je jeho autorem. Orientace v datech, týkajících se projektu, je velmi obtížná, např. v podadresáři DOC lze nalézt textový dokument, napsaný střídavě v českém, slovenském a anglickém jazyce. Ani v adresáři TRIE2, který by měl být relevantní vůči této práci, jsem neobjevil nic, co by připomínalo vývojovou dokumentaci; naopak zde lze nalézt řadu zbytečných záložních kopií různých souborů (.bak). Optimalizace, které jsou popsané v práci, nelze zřejmě volit přepínačem při volání programu, ale jen úpravou zdrojového kódu, což považuji za velmi nešťastné řešení.

V celkovém hodnocení je třeba ocenit, že v předložené práci autor prokázal jak jistou invenci pro návrh vlastního řešení zadaného problému, tak schopnost navržený postup implementovat a experimentálně vyhodnotit. Dosažené výsledky by si ovšem zasloužily o něco přesvědčivější zpracování a pečlivější dokumentaci. Přes výše uvedené výhrady se domnívám, že autor splnil hlavní cíl, uvedený v zadání tohoto tématu, a proto doporučuji, aby byla předložená práce přijata jako práce diplomová.

27. ledna 2009