Oponentský posudek disertační práce

Mgr. Jana Lánského *Syllable-based Compression*

Předložená práce se zabývá bezeztrátovou kompresí dat, konkrétně slabikovou kompresí textových a XML souborů.

Oblast komprese dat je dnes reprezentována především metodami ztrátové komprese obrazu, videa či zvuku, viz tolik populární formát MP3. To by mohlo budit dojem, že oblast bezeztrátové komprese dat stojí stranou tohoto boomu. Tento pohled je však velice zavádějící, protože výzkum bezeztrátových metod komprese dat se přesunul do oblasti vysoce specializovaných metod vyvinutých pro přesně daný účel - archivace dat s vysokým kompresním poměrem, komprese souborů na disku s dekompresí v reálném čase atd. Podobnou speciální oblastí je komprese velice malých souborů, kde typickým příkladem jsou SMS zprávy, či oblast komprese XML (HTML) souborů, kterou musí řešit každý internetový vyhledávač. Proto považuji zpracovávané téma za vysoce aktuální.

Struktura práce

Práce je členěna do osmi kapitol. První kapitolu tvoří úvod práce a vymezení jejího tématu. Druhá kapitola obsahuje základní teoretický aparát potřebný ke studiu dalšího textu. V kapitole 3 autor shrnuje současný stav poznání v dané oblasti. Kapitola 4 je teoretického rázu. Autor zde definuje základní pojmy, jako je slovo, slabika, algoritmy dekompozice textu na slabiky atd. Pátá kapitola přináší popis slabikové komprese textů, konkrétně metod LZWL a HuffSyllable, včetně dosažených experimentálních výsledků. V kapitole šesté se autor věnuje kompresi malých XML souborů, metodám XMillSyl a XMLSyl, opět včetně experimentálních výsledků. Kapitola 7 řeší kompresi rozsáhlých abeced, které je nutné přenášet spolu s komprimovanými daty při semiadaptivním způsobu komprese. Závěrečná, osmá kapitola se věnuje kompresi velkých XML dokumentů, s tím, že tyto dokumenty mohou být chybně formovány (angl. non-well formed). Klasickou ukázkou chybně formovaných XML dokumentů jsou HTML dokumenty. Výsledkem autorovy práce je projekt XBW.

Práce představuje vyspělou studii v oblasti komprese dat. Autor zásadním způsobem rozvinul stávající metody komprese dat adaptací slabikové abecedy, což lze považovat za nové, původní výsledky. Další významným vědeckým materiálem jsou v práci prezentované experimentální výsledky. Autor vlastními vědeckými výsledky značně přispěl k rozvoji poznání v dané oblasti informatiky.

Použité postupy práce jsou plně v souladu s obvyklou praxí v tomto vědním oboru. Formální stránka práce je na velice dobré úrovni. Výsledky disertační práce již byly prakticky použity v projektu Egothor.

Publikační činnost autora

Autor ve své práci uvádí seznam 21 publikací, z toho

- 17 ve sbornících konferencí (6 IEEE sborníků, 1 LNCS),
- 1 časopisecká publikace.

Obzvláště je třeba vyzvednout 5 příspěvků na konferenci *Data Compression Conference* konané každoročně v USA a která je komunitou zabývající se kompresí dat chápána jako nejlepší konference v oboru. Dostat se na tuto konferenci, byť s posterem, je úspěch zasluhující si zvláštní zmínku.

Připomínky

strana 14 Definice kompresního poměru je částečně matoucí. Bývá obvyklé kompresní poměr udávat v procentech. Vzhledem k tomu, že autor násobí daný poměr 8, dá se vytušit, že se bude jednat o veličinu nazývanou bits per character, bpc, viz např Salomon: Data Compression, 3. vydání, strana 10-11. Což, ale není explicitně uvedeno.

strana 34 „Our formalization has to be strong enough to be able to process also a document that can be in random binary form...“ Vzniká otázka, jak definovat slabiky či slova například v digitalizovaném zvuku?

strana 36 Co znamená znak & v definici 4.4? Konjunkci?

strana 39 Proč je stanovena maximální délka bloku samohlásek na 3? Empirie?

strana 40 „If α_i is a word decomposable into syllables, then the algorithm returns...“. Který algoritmus A nebo P?

strana 40 Definice 4.12. Jestliže je dekompoziční algoritmus P univerzální pro všechny jazyky L, pak nemůže existovat jazyk $L2$ pro který P není schopen provést dekompozici. V opačném případě by P nebyl univerzální. Asi by bylo dobré tuto definici přeformulovat.

strana 48 Věta „Sizes of sets...“ se vyskytuje v textu dvakrát bezprostředně za sebou a pokaždé uvádí jiné číselné údaje. Která varianta je správně?

strana 62 Tabulka 6.2 je dle mého názoru nepřehledná, názornější by byla orientace „na výšku“.

strana 63 Uvádíte, že ve většině případů stačí 170 položek pro značky XML dokumentu a cca 80 pozic v tabulce pro jména atributů. Pokud tyto počty nestačí, je možno rozšířit tabulku tak, aby používala dvoubajtové kódování. Jak víte, kolik pozic bude potřeba? Nebo se problém řeší adaptivně? Jak pak ale dekompresní algoritmus pozná, jak dlouhý kód pozice má očekávat?

strana 67 Údaje v tabulkách 6.3 a 6.4 jsem nebyl schopen pochopit. Uvádíte, že LZWL dosahuje dobrého kompresního poměru, zhruba dvou třetin metody XMill. Kde ty dvě třetiny jsou? Jako hlavní problém této tabulky vidím použití bpc jako měřítka kompresního poměru a použití veličiny nazvané zde kompresní faktor, která dává hodnoty číselně podobné kompresnímu poměru. Podle mého názoru by bylo dobré druhou jmenovanou veličinu udávat například procentuálně. Mimochodem jako kompresní faktor se ale většinou udává převrácená hodnota kompresního poměru, viz výše zmiňovaná kniha.

Nicméně výše uvedené připomínky považuji za spíše doporučení a postřehy co by se dalo na práci zlepšit.

Dále bych se rád zeptal na možnosti využití konceptu „spaceless words", mechanismu eliminace jisté slabiky (obdoba eliminace oběti) či slabik s jediným výskytem, viz disertace J. Dvorského.

Závěr

Jsem přesvědčen o tom, že předložená disertační práce přináší hodnotný příspěvek ke stavu vědeckého poznání v dané oblasti, autor prokázal své předpoklady k samostatné tvořivé práci a plným právem si zasluhuje udělení titulu doktora filozofie.

V Olomouci dne 14. listopadu 2008

[redacted]D.

[redacted]atiky

[redacted]rava