# Oponentský posudek diplomové práce
## Katsiaryny Chernik: Syllable-based compression of XML

Práce je věnována slabikové kompresi textových dat ve formátu XML. Jejím deklarovaným cílem bylo navrhnout algoritmus pro kompresi XML dokumentů, který bude pracovat nad abecedou slabik, navržené řešení implementovat a provést porovnání s existujícími metodami pro kompresi tohoto typu dat. Popis navrženého řešení spolu s nezbytným teoretickým pozadím a výsledky experimentálního vyhodnocení jsou obsahem textu práce. Podrobně komentované zdrojové kódy programů s pomocnými slovníky, uživatelskou dokumentací a soubory, jež byly předmětem v práci popsaných experimentů, tvoří její elektronickou přílohu.

Na textu práce, který se autorka rozhodla napsat v anglickém jazyce, oceňuji dle mého soudu kultivovaný výklad a srozumitelné vyjadřování, které je jen ojediněle narušeno drobnými gramatickými (str. 38: „it attempt" -> „it attempts", str. 40: „are metod" -> „are methods", str. 44: „reach morphology" -> „rich morphology", str. 37: nedokončená věta "... which was slightly.") či tiskovými chybami (str. 37 - 38: „XMLzwl" a „XMLlzwl", „CRF" zavedeno na str.31 a „CFR" použito na str. 38 zřejmě označují totéž).

Úvodní část textu je po stručném popisu formátu XML věnována výkladu dvou klasických metod, Huffmanově kódu a metodě LZW, které byly zřejmě vybrány jako reprezentanti entropických a slovníkových metod bezztrátové komprese dat. Vzhledem k tématu práce se výklad soustředí na varianty, které pracují nad abecedou slabik (či slov). Následuje poměrně reprezentativní přehled principů a metod používaných pro kompresi XML dokumentů. Podrobnější popis algoritmů XMill, XMLPPM, XGrind a Xpress je zde doplněn stručnou charakterizací sedmi dalších metod. Jádrem celé práce je další kapitola, v níž autorka navrhuje vlastní systém pro slabikovou kompresi XML dokumentů, pro který adaptuje dvě dříve popsané metody (LZWL, HufSyl) slabikové komprese textu.

Z mého pohledu nejzajímavější částí je následné porovnání algoritmů z hlediska dosahovaného kompresního poměru a interpretace výsledků. Autorka rozdělila dokumenty testovacího korpusu do tří tříd dle jejich charakteru (použitý jazyk a struktura), a ve třídách s jednoduchou strukturou a rozsáhlým textovým obsahem též do tří kategorií dle velikosti souborů. Na těchto datech je pak provedeno porovnání obou navržených metod slabikové komprese XML dokumentů (XMLzwl, XMLhuf) s jejich původními verzemi pro univerzální kompresi textu (LZWL, HufSyl) a dále s programem XMill, který byl zvolen jako referenční standart pro XML kompresi. XMill je zde použit v implicitním režimu, který využívá kompresní algoritmus gzip. V interpretaci výsledků se hovoří i o verzi algoritmů nad abecedou slov, v uvedených tabulkách (6.1-6.3, str. 38-39) jsem ale výsledky pro slovní kompresi neodhalil.

V další části experimentů autorka využila toho, že XMill lze zkombinovat v podstatě s libovolným kompresním algoritmem, a pro relativní porovnání účinnosti slabikové komprese testovala i variantu, kdy pro kompresi dat, které XMill extrahuje z XML dokumentu, je využita vždy jedna z obou zkoumaných verzí slabikové komprese, zatímco struktura dokumentu je komprimována - stejně jako v implicitním režimu - algoritmem gzip. Všechna uvedená hodnocení se bohužel nevyhnula běžnému nedostatku podobně zaměřených prací, soustředí se totiž výhradně na kompresní poměr, zatímco údaje o časové a prostorové náročnosti testovaných programů (podobně jako např. údaje o velikosti použitých slovníků častých slabik) zde chybí.

Analýza výsledků poskytuje cenné svědectví o chování slabikové komprese na XML dokumentech, klasifikovaných dle výše uvedených kritérii (velikost, jazyk, struktura). Pokud však jde o absolutní hodnocení účinnosti navržených metod ve srovnání s referenčním standardem, představovaným kombinací XMill + gzip, v některých specifických případech (divadelní hry v anglickém jazyce) sice slabiková komprese poskytla lepší kompresní poměr, celkové výsledky však dle mého soudu zatím nejsou natolik přesvědčivé, aby podpořily nasazení slabikových metod v této oblasti. V tomto ohledu bych doporučoval o něco střízlivější hodnocení výsledků nežli autorka, která v závěru práce označuje XMLSyl za velmi úspěšnou metodu komprese XML dokumentů. Je třeba brát v úvahu, je program XMill, s nímž byly navržené algoritmy porovnávány, je z existujících – v práci popsaných – XML kompresorů historicky nejstarší, přičemž jeho následníci dosahují – alespoň dle experimentů provedených svými autory – lepších kompresních poměrů (např. XMLPPM) nebo mají další výhody (např. umožňují vyhledávání bez dekomprese celého dokumentu). Autorka argumentuje tím, že program gzip, který je založen na kombinaci slovníkové metody LZ77 a Huffmanova kódu, dosahuje dle dříve publikovaných výsledků vyšší účinnosti nežli používané slabikové metody LZWL a HufSyl, a proto jsou i o něco horší kompresní poměry jejich adaptací XMLzwl a XMLhuf vlastně úspěchem. Nabízí se ovšem přirozená námitka, že je-li naším hlavním cílem zjistit, jaký efekt má použití slabikové komprese na XML dokumenty, bylo by vhodnější porovnávat systémy, jejichž kompresní jádro vychází z téže metody, tj. buďto jako standart použít kombinaci programu XMill s nějakou implementací LZW (např. klasický compress), nebo spíše navrhnout slabikovou variantu metody LZ77 a tu potom porovnávat s kombinací XMill+gzip.

Přes výše uvedené střízlivé hodnocení dosažených výsledků se domnívám, že jde o zdařilou práci, v níž autorka prokázala schopnost zvládnutí všech podstatných fází vlastního výzkumu: Schopnost nastudovat relevantní výsledky z literatury, dostatečnou invenci pro návrh originálního řešení, ale též schopnost navržené algoritmy implementovat, experimentálně vyhodnotit a na základě analýzy výsledků navrhnout možné směry dalšího vývoje. Je proto zcela na místě, že se část práce již podařilo publikovat ve sborníku konference. Protože autorčiny výsledky dle mého názoru splňují všechny požadavky, uvedené v pokynech pro vypracování tohoto tématu, mohu doporučit, aby byla předložená práce přijata jako práce diplomová.

15. května 2006

Tomáš Dvořák