Diplomová práce – posudek vedoucího

Katsiaryna Chernik: Syllable-based compression of XML

Předložená práce se věnuje kompresi textových dat ve formátu XML. Cílem práce bylo navrhnout a implementovat kompresní algoritmy využívající strukturu XML v kombinaci s algoritmy pro kompresi textu po slabikách. Obsahem této diplomové práce je teoretický základ, popis existujících a navržených algoritmů a experimentální část. Elektronickou přílohu pak tvoří zdrojové kódy programů, testovací korpus a výsledky experimentů. Přiložené kompresní programy fungují korektně (komprese i dekomprese) a to i na jiných než přiložených datech.

Práce je členěna do 8 kapitol. V úvodu je vysvětlena motivace celé práce. Druhá kapitola je věnovaná XML a jeho zpracování. V třetí kapitole jsou popsány existující obecné kompresní metody a kompresní metody pracující nad abecedou slabik. Další kapitola obsahuje popis metod specializovaných pro kompresi XML. Následuje kapitola o metodice provedených experimentů.

V šesté a sedmé kapitole jsou navrženy vlastní kompresní algoritmy XMLSyl a XMillSyl (každý ve dvou variantách pro LZWL a HufSyl). Algoritmy jsou podrobně popsány a otestovány na rozsáhlém množství dat v češtině i angličtině. Soubory jsou děleny podle velikosti na malé, střední a velké a dále podle typu na dokumenty s převahou textu a na dokumenty s převahou značek. Výsledky experimentů jsou vyhodnoceny a srovnány s existujícími algoritmy XMill, LZWL a HufSyll.

Kvalita textu je celkově dobrá, občas lze nalézt drobné překlepy. Například v českém abstraktu je kódý místo kódy, na straně 10 je syllable-base místo syllable-based.

Celkově by jsem chtěl ocenit, že autorka nastudovala poměrně nové výsledky z dvou různých směrů v oblasti bezztrátové komprese a vhodně navrhla jejich vzájemné zkombinování. Přínos práce je v rozšíření stávajících slabikových metod pro kompresi strukturovaných dat. Výsledné kompresní metody jsou na všech typech souboru lepší než klasické slabikové kompresní metody a na některých dokumentech jsou dokonce lepší než všechny ostatní srovnávané metody (tedy i XMill).

Práce nabízí dobrý základ pro další výzkum v této oblasti, například v kombinaci se slabikovou verzí algoritmu bzip2 nebo využitím informací z DTD dokumentu. Po malých úpravách lze algoritmy uzpůsobit také pro kompresi formátu HTML. Výsledky této práce byly úspěšně prezentovány na konferenci DATESO 2006.

Předloženou práci považuji po všech stránkách za práci splňující kritéria pro diplomové práce na MFF UK a doporučuji ji k obhajobě.

Praha, 11. 5. 2006

Mgr. Jan Lánský, KSI MFF UK