# Diplomová práce – posudek oponenta

## Václav Matouš: Efektivní ukládání html stránek

Předloženou práci lze rozdělit na dvě velké části. První část se zabývá předzpracováním HTML dokumentů do podoby vhodnější pro další kompresi pomocí metod gzip a bzip2. Druhá část se zabývá tvorbou úložiště pro webové stránky. Hlavní požadavek je kladen na rychlost přístupu a efektivní ukládání jednotlivých dokumentů (včetně verzování).

V první části práce autor provedl analýzu reálných HTML dokumentů a rozbor existujících norem pro jejich tvorbu. Na tomto základě stanovil nejčastěji se opakující značky, ze kterých vytvořil slovník. Tento slovník se použije v předzpracování dokumentu pro kompresi. Protože v klasickém HTML dokumentu není použito všech 256 znaků 8-bitové abecedy, tak nepoužité znaky nahradí značkami ze slovníku, které se v daném dokumentu vyskytly. Algoritmy nahrazování jsou optimalizovány na rychlost.

Autor se věnuje i částečně ztrátovému předzpracování, kdy zaniká informace o velikosti písmen použitých v názvu značek.

Předzpracované dokumenty jsou pak následně komprimovány metodami gzip a bzip2. V části experimentů autor porovnává výhodnost předzpracování dokumentu z hlediska úspory prostoru, zmíněn je i čas komprese, nicméně časy dekomprese zmíněny nejsou, stejně jako paměťové nároky. Výsledky jsou zpracovány ve velkém množství tabulek, kterým by jsem však vytkl nešťastně zvolené rozdělení testovacích souborů do kategorií podle velikosti. Autor zvolil kategorie do 10kb, 10-20kb, ..., 80-90 kb, 90-100kb, nad 100kb. Celých 76 % testovaných souborů se však nachází v jedné z prvních dvou kategorií, které měly být podle mého názoru více detailně členěny. Naopak pro větší soubory mohlo být kategorií méně. Rovněž by mohlo být zajímavé zkusit toto předzpracování použít i pro některou z metod pro kompresi malých souborů (Rein a kol. DCC06, Korodi a kol. DCC05), případně pro některou z textových kompresních metod.

Druhá část práce se zabývá tvorbou datového úložiště pro HTML soubory. Požadována byla podpora verzování souborů a rychlý přístup k poslední verzi dokumentu. Rovněž jednotlivé dokumenty byly komprimovány metodami z první části. Na závěr této části jsou opět rozsáhlé experimentální výsledky. Autor dosáhl jisté úspory obsazeného prostoru a zároveň čas přístupu k jednotlivým verzím dokumentů je přijatelný.

Domnívám se, že jde o velmi zdařilou práci, přestože v některých ohledech by šla ještě mírně vylepšit. Autor prokázal při zpracování tohoto tématu značnou míru vlastní iniciativy, práce byla proto právem přijata k publikování na konferenci ITAT 2007.

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

Praha, 29. 8. 2007

Mgr. Jan Lánský, KSI MFF UK