Posudek vedoucího diplomové práce
Přemysla Pašky: Kontextové modely pro statistickou kompresi dat

Klasické statistické metody bezztrátové komprese dat se obvykle popisují jako algoritmy tvořené dvěma relativně nezávislými fázemi - modelováním a kódováním. Zatímco problém kódování je považován za poměrně uspokojivě vyřešený, otázka tvorby statistického modelu dat je stále předmětem výzkumu. Deklarovaným cílem této práce bylo prozkoumat známé přístupy ke tvorbě modelů dat a poté se je pokusit zobecnit se zaměřením na maximální dosahovaný kompresní poměr.

Na textu práce, který je napsán v anglickém jazyce, oceňuji srozumitelné vyjadřování a velmi solidní jazykovou úroveň. V úvodu, věnovaném současnému stavu zkoumaného problému, si autor povšiml, že důvody podobné cílům této práce vedly k vypsání tzv. Hutterovy ceny, kde bylo zatím nejlepších výsledků dosaženo prostřednictvím metody PAQ. Autor tedy nejprve vysvětluje, proč se místo bitově orientovaného PAQu rozhodl vycházet z klasické metody PPM. V další kapitole popisuje problém nepřesných statistik v kontextech vyšších řádů a navrhuje vlastní řešení. Výsledkem je algoritmus pro konstrukci kontextového modelu dat, založený na datové struktuře, původně navržené S. Buntonovou (1996), kterou autor modifikoval pro potřeby reprezentace úplné historie dat. Další část práce popisuje pilotní implementaci navrženého řešení a jeho dvou modifikací: verzi s děděním informace podle D. Škarina (odst. 3.6.2) a varianty s váženou historií (kap. 4).

Experimentální vyhodnocení na souborech Calgarského korpusu bylo zaměřeno hlavně na dosahovaný kompresní poměr. Výsledky byly též porovnány se dvěma existujícími implementacemi metody PPM: klasickou variantou PPMD (Howard, 1993) a moderní verzí s heuristikami PPMII (Škarin, 2002). I když jako neúčinnější se ukázala metoda PPMII, prozkoumání trendu vývoje kompresního poměr v závislosti na maximální délce uvažovaného kontextu vedlo autora k několika zajímavým pozorováním. Zatímco existující varianty metody PPM ztrácejí účinnost přibližně na hranici kontextů délek 5 a 6, účinnost navržené metody se zlepšuje až po maximální testovanou délku 256. Tím se podařilo mírně zpochybnit obvyklou tezi o tom, že problém kontextů vyšších řádu je způsoben nedostatečnými statistikami. Podle mého soudu je škoda, že se autor při testování omezil na dnes již poněkud historický Calgarský korpus: testy nad rozsáhlejšími soubory či daty jiného typu (např. Slezský korpus) by možná mohly vést k ještě zajímavějším závěrům.

Práci lze samozřejmě vytknout i několik formálních nedostatků. Je sympatické, že se autor snaží v textu formulovat matematické věty a důkazy, takový přístup ovšem vyžaduje dle mého soudu poněkud přesnější vymezení používaných pojmů, nežli nabízí např. Definition 1 na str. 23. V části, věnované experimentálním výsledkům (odstavce 3.7. a 4.2), chybí přesná specifikace hardware, na němž byly testy prováděny. Program 3C, který realizuje v práci popsané ideje, je na přiloženém CD dodán s automaticky generovanou vývojovou dokumentací, ale případný uživatel by jistě ocenil i stručnou uživatelskou dokumentaci, která na CD chybí. Nejde ovšem o nijak zásadní nedostatky: Smysl Definice 1 je zřejmý z kontextu, testy kompresního poměru by použitým hardware a prostředím neměly být ovlivněny, a stručný popis spuštění programu lze nalézt v textu (odstavec C.2, str. 62).

V celkovém hodnocení práce bych ocenil, že se autorovi podařilo nastudovat a pochopit řadu netriviálních výsledků z daného oboru, navázat na tento výzkum vlastními idejemi, a navržené řešení implementovat a experimentálně vyhodnotit. Domnívám se, že obsah práce splňuje požadavky, uvedené v pokynech pro vypracování tohoto tématu, a proto doporučuji, aby byla posuzovaná práce přijata jako práce diplomová.

V Praze dne 18. září 2008