Oponentský posudek diplomové práce

Název DP: **Univerzální index textových dokumentů**
Diplomant: **Marek Švantner**

Obsah práce:

Předmětem diplomové práce (DP) je implementace full-textového indexu založeného na dynamickém invertovaném souboru. Po úvodní kapitole následuje přehled různých typů indexů pro použití v DIS. Třetí kapitola se věnuje návrhu univerzálního indexu - dynamického invertovaného souboru a v další kapitole autor analyticky rozebírá jeho efektivitu. Pátá kapitola rozšiřuje návrh indexu o transakční zpracování (index zajišťuje odolnost proti výpadkům) a v šesté kapitole jsou diskutovány různé kompresní techniky použitelné v rámci návrhu indexu. Sedmá kapitola popisuje implementační specifika a osmá diskutuje výsledky experimentů na reálných a syntetických kolekcích. Po závěrečné kapitole následuje ještě cca 30 stran příloh s technickými informacemi a CD.

Hodnocení:

Autor se zhostil úkolu velmi důkladně, návrh metody obsahuje jak teoretickou část (analýza aktualizačních nákladů), tak techničtější návrh, jako je komprese indexu a transakční zpracování. O důkladnosti svědčí také neobvyklý rozsah práce (120 stran). Rovněž implementační část naznačuje velký kus programátorské práce. Text práce je přehledný, srozumitelný a celkově je formálně na velmi dobré úrovni.

Jakkoliv byla práce vypracována důkladně, má jeden důležitý nedostatek. Tím je autorův mylný předpoklad (opakovaný na mnoha místech), že dosud neexistovaly jiné dynamické implementace invertovaného souboru ve smyslu efektivní aktualizace. To zkrátka a dobře není pravda, invertované soubory byly dávno dynamizovány a komerční implementace vyhledávačů typu Google pochopitelně musí používat techniky podobné těm uvedeným v této práci, jinak by vůbec nemohly fungovat. Na základě tohoto autorova předpokladu práce zcela postrádá rešerši o existujících alternativách a jejich odlišnostech od autorova přístupu. Totéž platí o uvedené literatuře, která je poměrně chudá a z větší části je redukována na skripta a přednášky MFF. Očekával bych citace např. těchto klasických článků:

- Cutting, D. and Pedersen, J. 1990. Optimization for dynamic inverted index maintenance. In Proceedings of the 13th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Brussels, Belgium, September 05 - 07, 1990). J. Vidick, Ed. SIGIR '90. ACM Press, New York, NY, 405-411.

- Tomasic, A., García-Molina, H., and Shoens, K. 1994. Incremental updates of inverted lists for text document retrieval. In Proceedings of the 1994 ACM SIGMOD international Conference on Management of Data (Minneapolis, Minnesota, United States, May 24 - 27, 1994). R. T. Snodgrass and M. Winslett, Eds. SIGMOD '94. ACM Press, New York, NY, 289-300

- Lim, Lipyeow and Wang, Min and Padmanabhan, Sriram and Vitter, Jeffrey Scott and Agarwal, Ramesh (2003) Dynamic Maintenance of Web Indexes Using Landmarks. In Proceedings International WWW Conference, Budapest, Hungary.

S uvedeným nedostatkem rovněž souvisí kapitola o experimentech, kde chybí jakékoliv srovnání s jinými systémy, např. systémem EGOTHOR, vyvíjeným dr. Galambošem na KSI (který je mimochodem rovněž dynamický). Experimenty tak vypovídají o efektivitě navrhovaného přístupu pouze nepřímo, sledováním různých nastavení parametrů.

Podrobnější připomínky, otázky, poznámky:

- V textu není zřetelně zdůrazněno, co přesně je převzato z práce panů Holuba a Míky, a co je autorův přínos. Nicméně předpokládám, že vlastním přínosem jsou kapitoly o transakčním zpracování a kompresi (a pochopitelně experimenty).

- Spíš než o kompresi bych hovořil o kódování, ale budiž. Nabízí se otázka, proč nebylo použito obyčejné statické Huffmanovo kódování – jistě by dopadlo lépe než Eliasovy kódy. Pokud nebylo zvoleno s ohledem na aktualizace, toto nebylo diskutováno v příslušné kapitole.

- V kapitole 2.1 tvrdíte, že vektorový model nelze efektivně implementovat pomocí invertovaných souborů. To není pravda, boolský a vektorový model se implementují prakticky stejně, rozdíl je v tom, že oproti uspořádání podle id dokumentů se v případě vektorového modelu uspořádává pomocí vah (uspořádávat lze podle čehokoliv, třeba PageRanku). Nicméně je pravda, že pokud se použijí tf·idf váhy, je aktualizace indexu obtížnější. Uspořádáním lze vektorový dotaz předčasně ukončit (relevance/podobnosti se postupně zvyšují stále pomaleji), tj. není třeba procházet pro daný term celý jeho seznam. Relevantní zdroje:

- V. N. Anh, O. de Kretser, and A. Moffat. Vector-space ranking with effective early termination. In Proceedings of the 24th annual international ACM SIGIR, pages 35–42. ACM Press, 2001.

- Moffat and J. Zobel. Fast ranking in limited space. In Proceedings of ICDE 94, pages 428–437. IEEE Computer Society, 1994.

- M. Persin. Document filtering for fast ranking. In Proceedings of the 17th annual international ACM SIGIR, pages 339–348. Springer-Verlag New York, Inc., 1994.

Závěr:

Práce splnila zadání a doporučuji ji k obhajobě.

V Praze dne 29. ledna 2007

RNDr. Tomáš Skopal, Ph.D.
oponent