Posudek vedoucího na diplomovou práci Igora Kulmana
Samoupravující seznamy

Tématem předložené práce je studium chování samoupravujících seznamů, což jsou datové struktury navržené pro rychlé vyhledávání za předpokladu, že některé z uložených prvků jsou hledány výrazně častěji než jiné. Úkolem diplomanta bylo vypracovat přehled známých algoritmů a porovnat je na základě publikovaných výsledků a vlastní experimentální studie. Práce svým obsahem toto zadání splňuje, ale zpracování mělo být lepší.

Úvodní teoretická část práce (kapitoly 1 -- 4) zavádí základní pojmy a popisuje známé samoupravující strategie. Autor vycházel zejména ze staršího přehledového článku z roku 1985, z novějších algoritmů připojil pouze randomizovaný algoritmus Timestamp z roku 1998. Současná problematika, např. vyšetřování modelů s lokalitou, je zde zmíněna jen velmi stručně odkazem na literaturu. Následující část práce (kapitoly 5 a 6) nejprve shrnuje výsledky předchozích experimentů, přičemž je čerpáno jak ze starší tak i ze současné literatury. Dále popisuje návrh vlastních experimentů diplomanta, včetně generování testovacích dat podle vybraných pravděpodobnostních rozdělení. Nejrozsáhlejší část (kapitola 7) představuje výsledky provedených testů formou komentovaných tabulek a grafů. V závěru je kromě shrnutí získaných poznatků uvedena i řada námětů pro další výzkum v této oblasti.

Těžiště práce spočívá v experimentech. Na rozdíl od předchozích experimentálních studií, v nichž byly algoritmy většinou testovány při vyhledávání v textových souborech, autor zvolil testy na číselných množinách. Tento poněkud umělý přístup umožňuje použití široké škály pravděpodobnostních rozdělení při generování dotazů. Pomocí nastavení jejich parametrů se tak dá nasimulovat nejrozmanitější průběh četností vyhledávaných prvků a zjistit, jak na tyto situace jednotlivé algoritmy reagují. V práci byla testována rychlost konvergence a rychlost vyhledávání ve stabilním stavu. Za nejzajímavější výsledek považuji zjištění, že chování algoritmů v situacích, kdy se vyhledávají jen některé prvky seznamu a rozdíly v jejich četnostech jsou výrazné, se v jistém smyslu podobá jejich chování v modelech s lokalitou.

Práce má bohužel řadu nedostatků, které snižují její úroveň. Konkrétně:
V úvodním přehledu algoritmů jsou až na výjimky všeobecně známé metody, bylo by dobré se podrobněji zmínit i o těch méně známých a hlavně o současných přístupech (aniž by všechny nutně musely být zařazeny do experimentální studie -- to by ani v diplomové práci nebylo z časových důvodů možné). Ale už vzhledem k tomu, že výsledky experimentů naznačují jistou podobnost s modely s lokalitou, mělo být o těchto modelech v práci více pojednáno.
Dalším nedostatkem jsou chyby a nepřesnosti v textu. Např. Věta 8 se omylem dostala k randomizované verzi algoritmu Timestamp, zatímco její tvrzení se vztahuje k deterministické verzi.
Samotný algoritmus Timestamp je popsán chybně a bohužel je také nesprávně naprogramován. To částečně znehodnocuje výsledky experimentů. Kromě toho mám pocit, že algoritmy jsou naprogramovány zbytečně složitě, což se může odrazit i v naměřených časech. Jednou ze zásad experimentální algoritmiky je použití co nejjednodušších progra-

movacích prostředků, tak aby byly vykonávány pokud možno pouze příkazy předepsané algoritmem a žádné jiné. Nevidím proto důvod k použití virtuálních objektů pro vcelku jednoduchou práci se spojovými seznamy. Rovněž odečítání (a následné ukládání) systémového času v okamžiku nalezení prvku v algoritmu Timestamp je zbytečné, protože čas, kdy je prvek vyhledáván, se dá jednoduše interpretovat jako jeho pořadí ve vyhledávací sekvenci.

Pro experimentální odhad kompetitivního faktoru měl být zvolen opatrnější název, protože se ve skutečnosti o kompetitivní faktor nejedná, pouze o jeho přiblížení – autor to sice ví a v textu to i na jednom místě uvádí, tabulky ale přesto nadepsal tímto nepřesným termínem.

U použitých pravděpodobnostních rozdělení většinou chybí diskuse o volbě parametrů. Proč je např. geometrické rozdělení použito s parametrem 0.013 a ne s nějakým jiným? Jak by se změnil průběh hustoty při jiné hodnotě parametru a co by to znamenalo pro vygenerované dotazy? Jak by se tato změna promítla do chování vyhledávacích algoritmů?

Domnívám se, že po odstranění těchto nedostatků by práce mohla být velmi slušná. V této podobě ji ale za diplomovou práci uznat nedoporučuji.

V Praze dne 23. května 2011