

# Posudek oponenta na diplomovou práci Romana Krejčíka

## *Relační modelování biologických dat*

Cílem práce bylo zmapovat existující metody pro ukládání hierarchických dat do databází, zejména pak reprezentovat grafové struktury biologických dat v relační databázi. Součástí práce mělo být i realizování vhodné implementace a a indexace takových struktur společně s experimenty na netriviálních kolekcích biologických dat.

V první části práce se p. Krejčík věnuje popisu obecných vlastností hierarchií, požadavkům na jejich uložení do DB a definuje i několik typů obvyklých dotazů. Ve střední – nejdělsí – části pak podrobně zkoumá známé metody uložení a detailně - zejména na úrovni SQL příkazů a tradičních indexů databáze - rozebírá, jak bude každá konkrétní metoda realizována. V závěrečné části se pak věnuje popisům vybraných kolekcí a praktickým testům.

### Pozitiva:

Práce je velmi přehledná, popsané metody jsou detailní a srozumitelné, jazyková úroveň je až na občasná překlepy výborná. Stejně tak příložené CD je dobře zpracováno a obsahuje i všechny potřebné soubory k případnému ověření provedených testů. Z textu je navíc patrné, že se autor problematice relačních databází věnuje a že svých znalostí umí příslušným způsobem využít.

### Negativa:

Jako jedno z významnějších negativ vidím faktické omezení hierarchických struktur na stromové grafy. Autor sice zadefinovává i další typy hierarchických dat, tyto vlastnosti však již dále příliš nerozebírá. Pokud ano, věnuje se zejména převodu takového typu dat na stromové, důsledky takového převodu však zmiňuje jen velmi stručně.

Pro srovnání jednotlivých metod používá pouze následující operace: nalezení otců/synů pro daný uzel, vyhledání uzlů v podstromu, zjištění cesty ke kořenu k danému uzlu, získání všech listů a nalezení společného předka. Diskuze, proč zrovna tyto operace, jaké jiné a z jakých důvodů by mohly být potřeba, chybí.

Z metod uložení stromových dat zmiňuje šest způsobů, které jsou skutečně velmi časté (a také často popisované v mnoha pracích) -- triviální stromové uložení, metoda s uložení cesty ke kořeni (dvě varianty), tabulku vazeb, vnořené množiny (využívající intervalové očíslování) a průchod do hloubky. Metody triviálního uložení, tabulky vazeb, prefixové cesty přechůdců i varianty intervalového očíslování jsou rozebrány ve skriptech MFF UK - XML Technologie, ze září 2006. Ve zmiňovaných skriptech (a článcích, na něž odkazují) jsou navíc zmiňovány i metody, které autor vůbec neuvádí (např. rozklad tabulek, využití vlastností složených typů, využití ORDPATH). Přidaná hodnota tedy spočívá zejména v podrobném rozepsání jednotlivých přístupů až na úroveň SQL dotazů (včetně dotazů pro aktualizaci hierarchie), protože do takového detailu většina publikací nezachází.


Obecná porovnání metod v teoretické části sice poskytují určitou představu o chování jednotlivých metod pro konkrétní typy dotazů, co to však znamená pro reálné zpracování dat není z práce patrné, protože na to nenavazují příslušné statistiky v závěrečné části práce. Formální porovnání složitosti ve střední části práce nám tuto představu také nedá - „hloubka stromu je obvyklém případě  $O(\log n)$ “ - toto tvrzení není nijak odůvodněno, ze str. 58-60 navíc vyplývá, že ani testovaná data nejsou příliš vyvážená. Odhady na tomto předpokladu (např. v kap 3.4 na str. 34) vycházejí samozřejmě příznivě, protože to vychází z ideálního stavu – odhadu pro úplný strom. Problematické je u využívání takového tvrzení pro výpočet délky maximálního identifikátoru – např. maximální hloubka podstromu 41 v testovacích datech NCBI je od ideálního stavu stavu cca hloubky 8 ( $=\log(329173)$ ) při základu 5.25, průměrného počtu synů) poměrně daleko. Autor si je tohoto rozporu nicméně vědom „to je sice příznivý odhad, ale o konkrétní velikosti pole nic neříká. Přesnou velikosti musíme zvolit podle charakteristiky konkrétních dat“.

Hlavní výsledek práce – chování jednotlivých metod na vybraných třech vzorcích biologických dat – je provedeno vždy nad náhodně vybranými 2000 uzly. Výsledky jsou však uvedeny jen k operacím cesta ke kořeni, dotaz na podstrom a pak dotaz na podstrom s řazením (str. 65 – 70). Pro další operace, které byly zmiňovány v teoretické části, nejsou žádné experimentální výsledky uvedeny.

Závěr:

Přes výše uvedené připomínky práce splňuje schválené zadání. Autor se dokázal detailně seznámit s poměrně obsáhlou problematikou a na vybraných metodách se pokusil ukázat jejich silné a slabé stránky. Bohužel právě experimentální část, která na rozdíl od teoretické není tak dobře popsána v jiných publikacích, zůstává v začátcích a i použitá metodologie pro provedení experimentů je spíše jen naznačena. Práci **doporučuji k obhajobě**, před obhajobou doporučuji klasifikaci **dobře**.

V Praze, 3. 9. 2007

  
oponent diplomové práce  
Mgr. Kamil Toman  
KSI MFF UK