Posudek oponenta diplomové práce

Název DP: **Indexing Linked Data**
Diplomant: **Andrei Conicov**

Obsah práce:

Předmětem diplomové práce (DP) byl návrh a implementace indexu pro RDF data, který by umožňoval efektivní vykonávání SPARQL dotazů. V první kapitole autor uvádí do problematiky Linked data, ve druhé pak sumarizuje základy modelu RDF. Třetí kapitola pojednává o existujících systémech pro indexaci RDF dat, přičemž v experimentech je autorova metoda srovnána se zde uvedenými systémy Sesame a Jena. Čtvrtá a pátá kapitola se věnují popisu samotného navrhovaného RDF indexu, rozsáhlá šestá kapitola pak obsahuje experimenty. Poslední kapitola práci shrnuje.

Hodnocení:

Autor odvedl velký kus práce – v podstatě musel implementovat nativní RDF index a SPARQL procesor nad tímto indexem. Rozsáhlost zadání bohužel ovlivnila návrh řešení, které by byl autor schopen v daném čase zvládnout, a tudíž výsledky experimentů nejsou přesvědčivé. Zde je ovšem na místě přiznat nerovný boj – softwarové dílo jednoho člověka v rámci diplomové práce vs. léta vyvíjené produkty (Sesame, Jena).

Organizace textu a výklad je místy neuspořádaný. Například to, že FullDoc index je originálním přínosem autora se dozvídáme postupně mezi řádky, a také nepřímo tím, že není citován jako již existující indexační struktura.

K analýze indexů na straně 45-46. Argumentace proti jednomu multidimenzionálnímu indexu ve prospěch tří B-stromů je zcestná. R-strom sice negarantuje logaritmickou složitost, tu ovšem negarantuje ani řešení pomocí tří B-stromů. Jednak může intervalový dotaz prostě vést na sekvenční přístup k polovině klíčů, jednak pro výsledné (velké) množiny je třeba spočítat průnik (a to pochopitelně na vnější paměti). Argumentace použitím R-stromu v geografických aplikacích je jistě pravdivá a R-strom nemá smysl používat pro vysokorozměrná data. Ve vašem případě je ale dimenze 3 a tudíž přesně odpovídá předpokladům R-stromu. Multidimenzionální B-strom i jeho blízký příbuzný UB-strom jsou také vhodnou variantou, stejně jako další prostorové indexy vycházející z vlastností B-stromu (X-strom). Výhodnost multidimenzionálních indexů pro nízkorozměrná data byla mnohokrát potvrzena (30 let stará technologie), je ovšem pravda, že do mainstreamových databází se i tak dostávají pozvolna.

Podrobnější hodnocení, připomínky a otázky:

V experimentech autor interpretoval rozdílnou rychlost indexace na dvou PC. Bohužel, obě PC byly téměř stejné platformy – více by pomohlo, pokud by jedno PC běželo na Linuxu. Ve Windows totiž nelze zcela kontrolovat systémovou cache a proto závěry ohledně dopadu implementované cache nejsou moc směrodatné, stejně tak efekt souborového systemu (proto taky systémy jako Oracle mají ve své enterprise konfiguraci vlastní filesystem).

Angličtina je slušná, trpí však řadou drobných neduhů, jako je např. chybná interpunkce či nepřesné odborné termíny (indexation, minimum binding rectangle).

Závěr:

Práce splnila zadání a doporučuji ji k obhajobě.

V Praze dne 13. května 2012

Doc. RNDr. Tomáš Skopal, Ph.D.
oponent