- jak souvisí váš „union“ s „merge“ používaným v popisu semantiky RDF v materiálech W3C?
- jak jste implementovali „!“?
- transformace stromu ze str. 26 na str.27 asi není v pořádku, protože uzel OPTIONAL je pak výše a má pod sebou všechny trojice a na začátku jen některé,
- tabulka na str. 28 je asi jen částečná, nevěřím že vám stačí např. T_1.subject=T_2.predicate a nepotřebujete např. T_1.subject=T_2.object; je T jméno pro triple tabulku nebo obecné?
- prosím vysvětlete co znamená na 64_7 „... všechny grafy, které odopovídají dotazu definujícímu index...“

Práce splnila cíl, je přínosem pro rozvoj experimentálních metod v informatice, splňuje nároky na diplomovou práci a doporučuji ji k obhajobě.

V Praze 16.5.2006

Prof. RNDr. Peter Vojtáš, DrSc.

Posudek na diplomovou práci

J. Dokulil. Dotazování nad RDF daty

Cílem práce bylo prostudovat současný stav dotazovacích jazyků nad RDF daty, provést jejich klasifikaci, srovnat jejich vyjadřovací schopnosti a implementovat jejich dostatečně silnou množinu konstrukcí pomocí (pokud možno) existujícího software.

Hned na začátek konstatuji, že cíle byly splněny - i když každý do jiné hloubky. Některé části jsou až příliš stručné, např. kap.2. Anglická terminologie základů RDF pochází z lingvistického subject predicate(verb) object a bylo by vhodnější to překládat českými analogiemi užívanými v lingvistice: podmět predikát(přísudek) předmět – vyhli bychom se pak kolizi s terminy z OOP (které vzniknou při pouhém počeštění anglických termínů).

Popis dotazovacího jazyka SPARQL je až příliš stručný (diplomand si ho vybral protože to je návrh konzorcia W3C v připomínkovém konání). Schází i víc formálnější model. Část o reprezentaci dat je jasná i když by si zasloužila více alternativ (ne jenom o výběru prostředí mezi DB2, Oracle, ...) ale i v modelování, protože to podle mne nejvíce ovlivnilo výsledky. Jenom kvůli modelování typů (str.32 „kvůli spůsobu jak je v nich nakládáno s typovým systémem RDF") přesunout všechno do dvou tabulek které pořád musím spojovat – myslím si že největší problém s efektivitou je skryt tady. Vyhodnocování SPARQL dotazů překladem do SQL je popsáno velice střídmě (s odkazy na literaturu) a určitě není „self-contained".

Těžiště a přínos práce je v implementaci (opět popsána velice zevrubně) a hlavně v testování. Je štěstím, že autor měl k dispozici velikou kolekci stohových dat (které jsou prakticky RDF daty) i když nad stohem není vytvořen obecný dotazovací jazyk ale jen proprietární systém používající SQL. Samotná kapitola o datech naznačuje že autor vzal testování seriozně a připravil si tady půdu pro kvalitní testy.

Nejvzácnější na celé práci je kapitola o dotazech a měření. Poprvé vidím seriozně navržen a zdokumentován experiment který snese i náročnější kriteria. To, že experiment je ovlivněn návrhem a nemožností ovládat optimalizátor Oraclu na věci nic nemění.

Samotné řešení navrženo v kapitole o indexech už postrádá formální model a pojem „index vytvořen na základě dotazu" je spíš heuristický. Autor asi neměl na mysli, že bychom v čase dotazu vytvářeli index. Navíc je v literatuře popsáno i horizontální ukládání RDF dat (což odpovídá jednoduchým indexům ve smyslu této práce) oproti vertikálnímu (stohovému) ukládání. To co autor navrhuje jako indexy, jsou zajímavé spůsoby ukládání a vzhledem k předešlému připomínají krok od relací k OLAP organizaci úložiště (i když tady nejde o dimenze v doménách ale o jakýsi rozklad (distribuovaných dat např. na webu) a znovu spojování objekt-atribut modelu).

Celkově je práce spíš heuristická, s funkční implementační částí a hlavně cennou experimentální částí. Je velmi dobrým startem do doktorandského studia, protože poskytuje úvodní testy do problematiky. Osobně bych přivítal srovnání výsledků s výsledky ze SeRQL (který data taky ukládá (v jistém režimu) do relační databáze) a překlad SPARQL do SeRQL by byl mnohem jednodušší (nebo s ručně přepsanými testovacími dotazy).

Podrobnější poznámky: