Posudek oponenta diplomové práce Daniela Crhy

Unified Querying of Multi-Model Data

Předložená práce se zabývá velmi atraktivním tématem: dotazováním v prostředí heterogenních informačních zdrojů. Heterogennost se netýká pouze datového obsahu, ale hlavně použití více datových modelů ve zdrojích dat. Formální přístup tvořící pozadí navrhovaného unifikovaného dotazování je teorie kategorií. Práce vychází z publikovaných prací týmu vedeného vedoucí práce doc. Mlýnkovou a využívá rovněž původní software vyvinutý týmem.

Autor svoji práci rozdělil na kapitoly týkající se datových modelů (kap. 1), zavedení aparátu kategorií a jeho použití pro dotazování (kap. 2) a stručného přehledu grafových datových modelů a dotazovacích jazyků (kap. 3). Z dotazovacích jazyků pak vybírá SPARQL jako základ pro vlastní návrh dotazovacího jazyka MMSQL (kap. 4) nad kategoriálním datovým modelem. Obecnější algoritmy pro jeho implementaci obsahuje kap. 5. Je velmi obsažná, místy málo strukturovaná, nabízející ovšem hlavní přínos autora práce, i když někdy s nejasnostmi. Naskýtá se např otázka, do jaké míry musí být části schématu popisující stejnou věc, kompatibilní. Co když jsou dvě stejné věci ve dvou schématech pojmenovány různě?

Těžiště praktičtější části práce tvoří kap. 6 s návrhem a prezentací MM-quecat, tj. vlastního projektu, který implementuje nejpodstatnější rysy MMSQL Implementace využívá již existující software MM-evocat určený pro modelování a vyhodnocování transformací model-kategorie. Z databázových modelů sloužících k integraci použil autor relační a dokumentový s databázemi PostgreSQL a MongoDB.

K diskusi:

s. 7 a další – škoda, že není jasně definován použitý E-R model. Je nejasný a se vzrůstajícím počtem stran se v něm objevují stále nové a nové rysy. Rovněž jeho použití místy nepřipomíná věty v přirozeném jazyce, jak je při použití konceptuálního modelu žádoucí. Např. Customer, Cart a Product na obr. 2.2. Cart není sloveso, konceptuálně tedy modelovaný typ vztahu nic neříká. Ještě horší to je u Order, Contact a Type. Jaký kontakt, jaký typ?

- v přehledu termínů populárních modelů se chybně ztotožňuje relace s tabulkou. Relace je množina n-tic, nikoliv tabulka.

s. 8 – Figure 1.1: u E-R schématu schází kardinality. Je ta kardinalita vztahu implicitně M:N? Náhledem do relační databáze se skoro zdá, že 1:N. U relačních dat není řečeno, jaké mají tabulky primární klíče. Hodnota null se v relačním modelu dat nevyskytuje. Ta je až v modelu tabulek SQL.

s. 9 – používaný E-R model je zvláštní. Je klíčový atribut a neklíčový atributu rozlišen plností resp. prázdností kroužku? Proč potom u Customer je v obr. 1.2 kroužek CustomerID prázdný? Totéž platí pro Product. Co je to typ entity Type a co je atribut (klíč?) Key? V JSON datech jsou atributy, které se vůbec nevyskytují v E-R schématu nalevo (např. cellphone, email).

s. 10: atributy _src, _tgt nejsou vysvětleny. E-R schéma popisuje kus reálného světa. Co popisují tyto atributy? A co Tag?

s. 11 – Obr. 1.4: kde jsou v E-R schématu zobrazeny Address a Phone?

- Obr. 1.5: Jaký je klíčový atribut typu entity Manual? E-R schéma spíše připomíná objektový přístup, kdy daný manuál má OID. Může mít Product více manuálů? Vzhledem k tomu že typ vztahu HasManual nemá explicitní kardinality, pak by taková situace mohla nastat. Co znamená u typu entity Product přerušovaná čára rámečku?

s. 13 – E-R schéma na obr. 2.2: spojování dvou atributů pomocí úsečky s plným kroužkem označuje další klíč typu entity? Co je však Tag? Dále: na obr. 2.1 se vyskytují v kolekci Order atributy cellphone a email, nikoliv však v E-R schématu na obr. 2.2.

s. 15, 17 – diagram kategorie schématu jednou neobsahuje kardinality, jednou ano.

s. 19: Požadavky na kategoriální dotazovací jazyk zmiňují čitelnost jazyka. Použití signatur v jazyku MMQL (viz kap. 4) ovšem jeho čitelnost nikterak nezvyšuje, spíše naopak.

s. 22: dotaz ve SPARQL vrací multimnožinu n-tic, nikoliv seznam.

s. 35: jaká je sémantika _: v obr. 4.4?

s. 42: krok 4 - jak systém pro vyhodnocení dotazu rozezná, zda má použít při přístupu k objektu pole nebo vnořený objekt. Jde o 2 různé dotazy.

s. 43: odkaz na nepublikovanou práci [32] nic nenapovídá o tom, zdali a kde bylo toto demo prezentováno.

s. 44: pro lepší porozumění jednotlivých fází v obr. 5.1 by bylo dobré vědět, co je DSL.

s. 87: co je to "semantic syntax"?

s. 88: který typ uživatele bude analyzovat různé plány dotazu?

s. 94: sémantika barevných obdélníků v obr. 8.6 (i 8.9, 8.11) je zcela nejasná.

Drobnosti: autor nepoužívá systematicky kurzivu pro definované pojmy (např. object, morpfism, composition na s. 13, naopak někde ano (např. static names, dynamic names na s. 18). Ovšem na s. 11 je pro několik pojmů dokonce použito tučné písmo.

s. 12: zkratka ER je zavedena, i když byla již dříve v textu používána.

s. 31: v tabulce chybí u PostgresSQL agregační funkce.

s. 32, 82, 87, 100 a další: odkazovat se na dosud nepublikované články není příliš přínosné. Je to jako na dům, který ještě nebyl kolaudován.

s. 88: text k obr. 7.1 není kompatibilní s textem 2. věty prvního odstavce kap. 7.2.

s. 89: text na obr. 7.2 není čitelný ani s brýlemi. Navíc téměř půl stránky neobsahuje nic.

s. 102: práce [5] – v čem byla publikována?

Obecněji: snad jen poněkud široký záběr vedl k tomu, že některé nesporně zajímavé myšlenky nejsou diskutovány podrobněji. Softwarové řešení nabízené v diplomové práci zřejmě předpokládá další rozvoj projektu. Možné je navázat a pokračovat v projektu, který již přesahuje rámec diplomové práce.

Závěr: Jde o velmi dobrou práci, která dokazuje hluboký zájem autora o věc, erudici při návrhu a vypracování složité softwarové architektury založené na zajímavém formálním základě. Otázky a problémy (viz výše) patří spíše do kategorie nejasností. Autor se pohybuje tvůrčím způsobem v modelové i dotazovací doméně, která je ve vývoji, jak dokumentují publikace autorského týmu z posledních let, který kategoriální styl uvedl ve světě databází do života. Práce je napsána v angličtině na velmi dobré jazykové úrovni.

V Praze dne 31.1.2023

Prof. RNDr. Jaroslav Pokorný, CSc.
KSI MFF UK

.