Diplomaterv-bírálat

Munka címe:	Hatékony gráflekérdezési technikák
Szerző:	Antal János Benjamin
Szak:	Mérnök informatikus MSc
Dátum:	2018. december 16.

Kitűzött feladatok

A diplomaterv két, kapcsolódó feladatot fog össze. Egyrészt arra keresi a választ, hogy milyen hatékony módszerrel lehet inkrementálisan karbantartani az egyes lekérdezések eredményét gráfadatbázisokban, másrészt ilyen gráfadatbázisok teljesítménykiértékelését tűzi ki célul egy benchmark keretrendszer felhasználásával és továbbfejlesztésével.

Szerkezet, felépítés

A dolgozat bevezetőjében a jelölt röviden ismerteti az adatok gráfként való tárolásának néhány lehetséges előnyét. Célként tűzi ki, hogy összemérje a már régóta alkalmazott klasszikus relációs adatmodellel dolgozó, és az adatbázisok világában még mindig újnak számító gráf adatmodellre alapuló adatbázis-kezelők alkalmazhatóságát azokra az esetekre, amikor gráf jellegű lekérdezéseket kell futtatni a tárolt adatokon.

A második fejezet tartalmazza a szükséges háttérismereteket. Kezdetnek bemutat egy egyszerű esettanulmányt, melyet később a különböző adatmodellek és lekérdezőnyelvek közötti különbségek illusztrálására használ. Ezt követően három adatmodell alapvető matematikai definícióit adja meg és részletezi jellemzőiket. Ezek az adatmodellek a tulajdonsággráf, szemantikus gráf és relációs adatmodell. A továbbiakban kitér a relációs lekérdezőnyelvekre, bemutatva a leggyakoribb operátorokat, valamint a kiválasztott nyelveket Cypher, Gremlin, SPARQL és SQL, és összefoglalja képességeiket, valamint összehasonlítja erősségeiket. A 2.4. alfejezet konkrét adatbázisokat sorol fel, röviden összegezve jellemzőiket. A második fejezet utolsó szakasza a dolgozatban leírt egyik újítás, az inkrementális nézetkarbantartás hátterének összegzésével foglalkozik, ami egy fontos kiegészítés a dolgozatot megelőző korábbi TDK munkához képest. Ismerteti az inkrementális nézetkarbantartás alapvető technikáit (algebrai, procedurális), valamint részletesen foglalkozik a differenciális adatfolyam számítási modellel. Zárásképp felsorol néhány technológiát, melyek közül a modellezést támogató megoldásokról itt még nem látszik világosan, hogy is kapcsolódnak a 2.5. alfejezet korábbi részeihez.

A teljesítménymérési keretrendszert a harmadik fejezet tárgyalja, és a jelölt itt mutatja be annak SPARQL és Cypher nyelven írt kiterjesztését. Kezdésnek összegzi az LDBC Social Network Benchmark Business Intelligence és Interactive terhelési profiljait, valamint a benchmark keretrendszer működését. Ezt követően beszél a keretrendszer kibővítése kapcsán végzett munkáról, melyben profilonként részletezi a munka során felmerülő kihívásokat és megoldásokat.

A negyedik fejezet a Transformation Tool Contest kapcsán végzett munkáról szól, melyben az inkrementális nézetkarbantartás problémáját címzi meg a Közösségi Hálózat feladat megoldásával.

A dolgozat másik hangsúlyos pontját az 5. Kiértékelés fejezet mutatja be. Az 5.1. alfejezetben a jelölt arra keresi a választ az LDBC SNB segítségével, hogy melyek azok az esetek, amikor az újnak számító gráfadatbázisok teljesítményben felülmúlják a régóta fejlesztett relációs adatbázisokat. A megtervezett mérési scenáriók valamint mérési elrendezés alapos bemutatása után ismerteti a kiértékelések eredményét a két terhelési profil esetén. A lekérdezések futási idejeit összesítő 5.1. ábrán érdemes lett volna az egyes részdiagramok y tengely feliratait egyeztetni, mivel a mérési eredmények egy nagyságrendbe esnek, továbbá milliszekundum helyett másodpercben megadni a mért értékeket. Az eredményeket a jelölt megfelelő részletességgel elemzi, értékeli. Az 5.2. alfejezetben pedig a differenciális adatfolyamok teljesítménymérését nézi a bemutatott Transformation Tool Contest Q1 és Q2 lekérdezésének segítségével, és hasonlítja össze a saját implementációját létező megoldásokkal.

A hatodik fejezetben a jelölt bemutatja a szakterület kapcsolódó munkáit, mely során beszél a leggyakoribb gráflekérdező keretrendszerekről, létező gráflekérdezési benchmarkokról és természetes illesztések optimalizálásáról.

A záró fejezetben a hallgató tömören felsorolja saját eredményeit, mely a korábbi, párban készített TDK munka miatt különösen fontos. Emellett jövőbeli tervnek a teljesítménymérési keretrendszer bővítését hozza fel, melynek során további Cypher támogatást, valamint az LDBC SNB mindkét profilját tovább szeretné fejleszteni. Ugyancsak jövőbeli terv a TTC Közösségi Hálózat feladatának megoldása, amihez a differenciális adatfolyam modellt tervezi használni.

Értékelés

A dolgozat olvasmányos, a használt alapfogalmak megfelelő mértékben bemutatásra kerülnek. Értelemzavaró elírások nincsenek, apróbb helyesírási hibák előfordulnak (pl. absztrakt második bekezdése: „gráfadbázis-kezelő”, 10. oldalon a 2. sorban: „illesztész”, kicsit lejjebb ugyanezen az oldalon a 2.3.1.2. fejezetben: „A r × s kifejezés egy n sorból álló r és egy m sorból álló s relációból [...]”, 18. oldalon „használ NULL értékeket”), melyek jelentős része egy helyesírás-ellenőrző futtatásával könnyen kiszűrhető lett volna. Továbbá a magyar és angol szavak keveredése kerülendő, ahol van elfogadott magyar megfelelő (pl. a 19. oldalon említett „batch-es számításokat” kötegelt számításoknak nevezzük). Irodalomjegyzék, hivatkozások megfelelőek, leszámítva egy önhivatkozást a jelölt korábbi TDK munkájára, hiszen jelentős része a dolgozatnak onnan van átvéve és ezek a részek nem új munkák.

A hallgató javarészt teljesítette a feladatkiírásban foglaltakat, sőt, van ahol a kitűzött célokat meg is haladta: mintegy szorgalmi feladatként megoldotta a 2018-as TTC Közösségi Hálózat feladatának Q1 és Q2 lekérdezését. Ezzel szemben a dolgozat nem fektet elég hangsúlyt a természetes illesztések témakörére. Bár ezek optimalizálásával kapcsolatosan végzett irodalomkutatást (6.3. fejezet), de lekérdezésoptimalizáló prototípust nem készített, helyette az inkrementális nézetkarbantartásra és ennek technikáira került hangsúly. Továbbá elmondható, hogy teljesítménykiértékeléssel kapcsolatos feladatait maximálisan teljesítette.

Mindezeket figyelembe véve a dolgozat színvonala és az elvégzett munka minősége több, mint megfelelő. A jelölt az elvégzett munkával tanúbizonyságot tett arra, hogy képes önálló mérnöki feladatmegoldásra.

Kérdések a jelölthöz

- A háttérismeretek fejezetben szerepel, hogy „[Az inkrementális nézetkarbantartás] két csoportját különböztetjük meg: az algebrai és a procedurális megközelítéseket”. A differenciális adatfolyamokat használó technikát hova sorolja, és miért?
- A 3.2.2.3. szakaszban bemutatja a példa Cypher lekérdezés Gremlinre fordított változatát. Ugyanitt írja, hogy ezek az automatikusan előállított Gremlin lekérdezések minimum szuboptimálisak, és a felmerülő technikai problémák miatt nem sikerült ezen a téren a dolgozat elkészültéig jobb eredményt elérni. Ugyanakkor a 2.4. fejezet technológiai táblázata alapján felmerül a kérdés, hogy pillanatnyilag a gyakorlatban van-e valós igény a Cypher for Gremlin támogatására, hiszen mindkettő tulajdonsággráfok feletti lekérdezésekre használható? Ha igen, mit tudna felhozni példának?
- Az 5.2. ábra alapján az is megfigyelhető, hogy a 2. és 10. (és esetleg még a 11.) lekérdezések jól látható teljesítménybeli sorrendet állítanak fel az egyes rendszerek között. Ki tud-e valami közös jellemzőt emelni ezen lekérdezések esetén, amik ezt a jelenséget okozhatják?

Montreál, 2019. január 3.

Búr Márton

okl. mérnök informatikus