Posudek oponenta na diplomovou práci
Vojtěcha Šípka

Comparison of Approaches for Querying in Chemical Compounds

Vyhledávání v chemických databázích využívá dotazování nad grafickými databázemi. Jde o aktuální téma, třebaže implementací odpovídajícího software existuje mnoho a to jak na komerční tak výzkumné úrovni. Cílem práce bylo porovnat tyto techniky, pokud možno v jednotném prostředí, což je obtížné právě v důsledku roztříštěnosti hardwarových a softwarových prostředí v pozadí.

Práce obsahuje úvod, čtyři kapitoly a závěr. V úvodní kapitole autor předkládá úvod do problematiky, zejména pak praktické motivace. Cíle práce, tj. dále studované techniky a strategii vlastního řešení jsou velmi stručně popsány v úvodu. Kapitola 1 obsahuje základní formální definice grafových pojmů a stručnou informaci o dotazování na podgrafy. Kapitola by mohla být podrobnější, např. orientovaný acyklický graf (viz s. 14) zde definován není, podobně ani sufixový strom (s. 8) nebo prefixový strom (s. 10). Čtenář se zde rovněž nepřímo dozví, že uvažované grafové databáze budou zřejmě množiny grafů.

V kapitole 2., sekci 2.2, diplomant popisuje některé indexační metody relevantní pro daný problém, tj. dotazování na podgrafy. Popis metod není formální, což vede k tomu, že je místy nejasný. Např. GraphGrep na obr. 2.1 neobsahuje žádná označení hran. Objevují se zde další nedefinované pojmy/zkratky, např. DFS nebo „size-increasing support function“ (s. 10). Porozumění dalším indexačním metodám vyžaduje navíc alespoň základní znalost konkrétních chemických pojmů použitých pro popis chemických struktur, což bohužel práce neobsahuje. Zmíněny jsou rovněž existující výsledky porovnání metod na různých benchmarcích. Sekce 2.3 popisuje dotazování na podgrafy ve dvou typech databázích – relačních a grafových. Bohužel, tím že není dáno schéma odpovídající relace reprezentující graf, je obtížné prezentovat jakékoliv SQL dotazy. Sekce 2.4. uvádí tři příklady komerčních řešení daného problému.

Kapitola 3 je věnována experimentům. Jsou zde formulovány hypotézy týkající se indexačních metod, typů dotazů a benchmarků. Pojmy velký a malý dotaz nejsou nikterak specifikovány. Bylo poněkud iluzorní se domnívat, že od autorů lze získat implementaci. To není reálné a je naopak na diplomantovi, aby metody implementoval v jednotném prostředí s podobnými technickými předpoklady. Jinak je těžké cokoliv srovnávat. Kapitola je velmi stručná, o tabulce v SQL se dozvíme pouze jména sloupců, ani její název, natožpak typy sloupců nejsou explicitně vyjádřeny. Žádné indexy nebyly použity? Zajímavější je použití PGQL. Mimochodem, je něčím podobný SQL, ale klausulemi MATCH rovněž jazyku Cypher.

Detailněji jsou experimentální výsledky popsány v kapitole 4 – zřejmě nejhodnotnější části práce. Výsledky jsou nakonec použity k potvrzení či vyvrácení hypotéz.

Nejlepší částí práce je závěr shrnující přehledně použité metody a získané výsledky.

V kritickém pohledu na práci lze objevit jeden základní nedostatek: přílišná stručnost některých partií (viz výše). Na druhé straně lze vyzdvihnout dobrou angličtinu práce.

Závěr: Na práci je vidět, že byla dokončována ve spěchu. Diplomant se nicméně zhostil úkolu, pronikl do problematiky a realizoval potřebný software. Zkušeností z implementace mohou být inspirující zejména pro další experimenty s dalšími metodami a dalšími sadami dat v aplikačním prostředí. Doporučuji práci přijmout za práci diplomovou.

V Praze dne 05. 06. 2019

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