Posudek oponenta diplomové práce

Název DP: **Deduplikční metody v databázích**
Diplomant: **Petr Vávra**

Obsah práce:

Předmětem diplomové práce (DP) bylo studium a srovnání metod odhalování duplicit v databázích, implementace vybrané deduplikační metody a diskuse jejích výsledků. Po úvodních kapitolách autor v kapitole 3 uvádí čtenáře do problematiky unifikace (resp. deduplikace) datových záznamů a ve čtvrté kapitole se věnuje rešerši jednotlivých deduplikačních metod. V páté kapitole následuje popis zvolené metody (Robustní deduplikační metoda), kterou autor rovněž implementoval. Předposlední kapitola obsahuje jak kvalitativní srovnání metod deduplikace, tak srovnání zvolené metody s komerčním produktem DataFlux. V závěru autor shrnuje práci a nastiňuje možnosti pokračování ve vývoji metody.

Hodnocení:

Práce je kvalitně zpracovaná, formálně a strukturou na dobré úrovni. Přestože je problematika značně složitá a v detailech vyžaduje netriviální matematický aparát, autor zbytečně nezabíhá do formalismů, neboť to ani na takto vysoké úrovni (metody a metodologie práce s relačními daty) není nutné. Rešerše deduplikačních metod i vybraná metoda jsou pěkně popsané, až na níže zmíněné připomínky. K experimentům mám vážnější připomínku, rovněž níže zmíněnou. Celkově se mi ale práce líbí.

Podrobnější hodnocení, připomínky a otázky:

1) V práci nepadlo jediné slovo o komplementárním pohledu na konzistenci databáze, tj. na redundanci ve smyslu klasického databázového návrhu (normální formy a funkční závislosti). Je zřejmé, že deduplikace se konzistenci věnuje zdola nahoru, tj. že se zajímá o data, nikoliv o schémata, kdežto normalizace se děje právě pouze na schématech (ještě před plněním databáze daty). Zde by se ale oba pohledy mohly vhodně kombinovat a autorem uváděné příklady to potvrzují. Např. v sekci 5.2.1 autor uvádí příklad s PSČ, kdy duplicita PSČ naprosto nic nevypovídá o duplicitě celého záznamu a tedy by neměla být zohledňována. To je ovšem zmíněno u tohoto příkladu ad-hoc (český čtenář ví, co je PSČ), systematičtější řešení by bylo zahrnout do deduplikačních metod funkční závislosti, kde by se např. lehce zjistilo, že Město, Ulice → PSČ. Další příklad – Tab 5.1 uvádí příklad s hudebními skladbami. Zde by ze znalosti závislosti Název písně (+ rok, země, atd.) → Skupina např. plynulo, že ve zmíněných vahách (str. 32) bude mít duplicita v atributu Skupina menší hodnotu než v atributu Název písně. Jinými slovy zde poukazuji na závažnou skutečnost, totiž, že ne všechny "duplicity" je žádoucí odstranit. I z hlediska relačního návrhu se snažíme vyhnout redundancím v databázi (tj. duplicitám), nicméně toho nedosáhneme nikdy absolutně – duplicity jsou v minimální míře nutné (referenční integrita, a vůbec modelování atributových vztahů v databázi). Deduplikační metody by měly umět rozlišit mezi duplicitami vyplývajícími z relačního návrhu (z množiny funkčních závislostí, např. Tab 5.1) a duplicitami zavlečenými jako chyba v datech.

2) Zvolená metoda je poměrně zajímavá, nicméně jí hrozí nekorektní chování v případě, že se neodhadne obvyklá kardinalita duplicity (velikost shluku). Jestliže je počet výskytů duplicity v databázi větší než k + 1, duplicita se vůbec nenalezne. Extremní případ je tabulka 100 totožných řádků a tabulka 100 náhodně vyplněných řádků – z hlediska metody dopadnou obě tabulky stejně, tj. žádná duplicita. Příliš mnoho duplicit se zvrhne na 0 duplicit.

3) Pozor na terminologii – termín metrika autor používá v obou obvyklých smyslech – jak obecně jako míru ohodnocující nějaké skutečnosti reálného světa, tak v matematickém slova smyslu metrickou vzdálenost (např. editační), která musí splňovat axiomy metriky. V práci se oba smysly tohoto pojmu nebezpečně prolínají (navíc, např. vzdálenost Spedis jistě nebude matematická metrika, vzhledem k normování a dalším operacím). Vlastnosti metriky (metrické axiomy) jsou přitom v práci velmi důležité, neboť většina algoritmů vzdálenostních deduplikačních metod spoléhá např. na symetrii a tranzitivitu (tedy např. trojúhelníkovou nerovnost). Tranzitivní uzávěr přes vzdálenosti by byl nesmysl, pokud by vzdálenost nebyla nějak tranzitivně "zkrocená", např. trojúhelníkovou nerovností.

4) Zmíněná efektivita algoritmů nad DBMS by se dala zvýšit tím, že by se místo indexů v databázi použil metrický index (za předpokladu, že použitá vzdálenost je metrika). Bohužel, v současných DBMS nejsou sofistikované metrické indexy stále přítomny.

5) Tranzitivita a reflexivita jsou definovány pouze pro binarní relace na množině, tj. autor tyto pojmy bez potřebného zobecnění používá nekorektně na databázových relacích!

6) Ačkoliv práce obsahuje jakési experimentální kvazi-srovnání (metoda Compare), nemohu souhlasit s tvrzením, že metody nelze srovnat ve smyslu přesnosti a úplnosti (vzhledem k subjektivně ohodnocené kolekci dat). To lze dokonce přímo z výstupů metody Compare. Autor argumentuje tím, že duplicity nelze v databázi jednoznačně určit, a proto by bylo srovnání ve smyslu přesnosti a úplnosti zavádějící. Tím ale v podstatě říká, že vůbec nemá smysl kvantitavně srovnávat dvě metody. To je ovšem alibismus (anebo byl pravý důvod vyhnutí se přímého srovnání s komerčním produktem?), protože přesnost a úplnost se většinou vztahuje právě k subjektivně ohodnocené kolekci. Např. konference TREC a její kolekce obsahují takové subjektivní anotace (třebaže pro jiné účely – information retrieval). Je jistě možné, že subjektivita takových anotací, v našem případě identifikace duplicit, vede ke zkreslení, protože jiný subjekt by ohodnocoval jinak. To nicméně nebrání použití více kolekcí, většinou by však měl stačit větší počet anotovaných duplicit v rámci jedné kolekce, kdy naměřená přesnost a úplnost bude v průměru spravedlivá ke všem testovaným metodám, nikoliv pouze k jediné (bernou mincí bude relativní srovnání, nikoliv absolutní). Zkrátka, metody je třeba srovnat tak, aby si uživatel (klient či šéf) mohl vybrat. Anebo na dotaz "který produkt je pro naše konkrétní data lepší" šéfovi odpovíte "no to záleží, radši koupíme všech 5 produktů, bude to sice stát 5x tolik, ale pro jistotu"? ☺

Závěr:

Práce splnila zadání a doporučuji ji k obhajobě.

V Praze dne 22. srpna 2010

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