Conditional Functional Dependencies for Data Cleaning

This paper's main contribution is using conditional functional dependencies (CFD) for data cleaning in database systems. The authors developed new methods for CFD violation detection in SQL with multiple constraints. They tried to use CFD with multiple constraints to clean the data with a reasonable performance. The introduction section explains why data cleaning is important, what are the CFDs and the difference between FDs and CFDs. The future plan of the authors is using CFDs to repair the data.

This paper has a great introduction section. It's simple and explains the problem very well also, the examples are pretty good. The other sections also have some examples about the topic. Especially FD8 example with a proof explains some of the FD rules. Also, after the FD rule table, authors describe each of the FDs individually for better understanding. They also use previous axioms like Armstrong Axiom to reinforce their theorems. Finally, last experimental study reinforces the all study.

It was a hard paper to understand. When I first looked at the paper, I saw lots of different symbols and they scared me. Thus, the paper isn't for people who are not familiar with the topic. The other aspect I don't like is the Figure usage. Most of the sections the authors mention the previous Figures. That's okay for most of the time but when I read this paper, I have to scroll up and down a lot of times. FD descriptions are decent but examples for them would benefit readers like me. Also, CFD queries seem very long and complicated and they mentioned that they used sql "case" query for some of the CFD queries. They also refer case isn't available for every DBMS (only DB2) on the market.

Like I mentioned in the weakness part, the authors should write a sentence about the previous Figures/Tables instead of mentioning them. With that, readers don't have to go up and check the Figures and its values to understand the topic. Also, at least some basic examples about FD rules would be nice. If the reader doesn't have a deep knowledge about the topic, these examples help the readers to figure out some of the rules and their explanations. The last section of the paper refers data repairing with CFD. I think the authors should have allocated more sections about CFD data repairing.

2019-01-20

Baran Kaya, kayab@mcmaster.ca, 400284996