Swoosh: a generic approach to entity resolution

The paper introduces 3 different algorithms for Entity Resolution (ER). These algorithms consist of 2 main parts: matching the records and merging them. The authors behave the match algorithms as a black box and try to increase this black-box's performance with fewer record comparisons. 2 of the algorithms uses ICAR properties which are defined in the paper. If these properties hold in records, swoosh algorithms' performance increases.

Introduction of the paper is great and summarizes most of the concepts. Definitions and examples are understandable and great for new readers. The authors present 3 different algorithms and some properties about them. These properties/rules are similar to the set rules thus, readers who are familiar to the mathematics get the concepts easily. The more the algorithm requires features, the better the algorithm's performance. That is why F-Swoosh is faster but requires features to compute. Also, not comparing the same records is cleaver way to increase the performance of the algorithm. Another good thing is they used real data for their experiments and evaluations. Finally, the paper's format is great. Especially, the proofs location. That way, only the readers who are interested with the subject can read them and others doesn't have to look at them if they don't want to.

There are a few weaknesses that I found while reading this paper. One of them is ICAR properties. Properties themselves are pretty good; however, I couldn't find any definitions about what does "ICAR" means in the paper. Also, I couldn't understand the difference between domination and merge domination. I think the authors could have explained it better with more examples. 3 algorithms are great but I don't think the paper explains when to use each one. They mentioned that when the new data added into the system F-Swoosh handles it very well. Yet, they didn't mention the same thing for G-Swoosh and R-Swoosh. Also, the example for F-Swoosh where it merges ("John Doe") names, what happens when the 2 names are the same but their other information are completely different. There maybe more than one person who share the same name.

They mentioned the brute force entity resolution method BFA in the beginning of the paper. I understand that their methods are completely different than the similar algorithms but I would like to see a performance comparison for different algorithms. All the graphs and evaluation part consist of 3 algorithms comparison. Also, there was a visual for brute force ER and I would like to see similar visuals for the new 3 algorithms as well. The authors mentioned very little about when to use which algorithm. They should have explained each algorithm and their use cases with examples (preferably real world examples).

2019-01-27

Baran Kaya, kayab@mcmaster.ca, 400284996