CAS 764 – Final Project Ideas

For the final project of the Advanced Topics in Data Management course, I would like to work on the duplicate detection in databases. In one of the lectures, you mentioned there are different tools for duplicate detection and data cleaning. I would like to compare some of the open source and commercial data profiling tools for their duplicate detection performance (accuracy and speed). However, I don't know how should I compare them. Also, one quick Google search about data profiling tools gave me these tools: Open source tools: Quadient DataCleaner, Aggregate Profiler, Talend Open Studio and for Commercial tools: IBM InfoSphere Information Analyzer, Data Profiling in Informatica, Oracle Enterprise Data Quality, SAS DataFlux. I am not sure if I can get/use these commercial tools for the project but I think some of them has demo or limited time licence that I can use. I would like to use data profiling for duplicate detection instead of data mining but if the scope of the project wouldn't be enough, we can extend it to data mining tools. I found similar papers about different data profiling techniques (distance methods, token based methods, …) and overall surveys about them. I would like to hear your ideas about which methods and datasets should I use for comparing data profiling duplicate detection performance.

2019-01-20

Baran Kaya, kayab@mcmaster.ca, 400284996