Data Curation at Scale: The Data Tamer System

Michael Stonebraker  Daniel Bruckner  Ihab F. Ilyas  George Beskales
MIT  UC Berkeley  QCRI  QCRI
Mitch Cherniack  Stan Zdonik  Alexander Pagan  Shan Xu
Brandeis University  Brown University  MIT  Verisk Analytics

1. Summary

Data Tamer is a curation (data integration) system with a sequence of data sources as an input, and it outputs a composite which is constructed over time. Their method includes Scalability, Data Cleaning, it is easier to use for non-programmers, and new data can be incrementally added. The authors tried their system with 3 real curation problems and it showed a significant reduction in cost.

2. Positive/Strong Points

2.1) They showed that data curation is a problem relevant to several different areas, such as Computer Science, Biology, and Health Services.

2.2) The integration of dictionaries (for example, a database with all the cities in the US and the state they are in) is a good implementation for examining the validity of the database.

2.3) Data Tamer learns with time, so that at the beginning it asks the user many questions but as time passes it asks less and less.

3. Negative/Weak Points

3.1) Putting the data in a key-features format that is readable for Data Tamer might be very costly. This part of their system requiring a "upstream wrapper" is not yet optimised.

3.2) The complexity of the algorithm used for Data Tamer is quadratic, which can be very costly for big data bases.

3.3) Data cleaning often relies on outlier detection, and they are not necessarily errors. However, they are treated as missing values.

3.4) When comparing Data Tamer with other method, they only have an estimated amount of repeated values. Data Tamer finds more of these duplicates, but it is difficult to know if there were actually that amount of duplicates. It could be that the data was selected so that the number of estimated duplicates would be closer to the number of duplicates found by Data Tamer.

4. Research Questions and Points for Discussion

4.1) It would be possible to identify the relationships between different entities (for example, entities into classes), which, although it might increase the cost, I think it would greatly increase the quality of the integration.
4.2) It seems to me that they are very much concerned about developing Data Timer to do the curation without really being concerned about optimization of their methods. This can be ok for early stages of the development, but will be a problem when big databases are implemented.

4.3) It could be possible to also add some statistical features to the data integration, which could give a useful description of the database and would help in its analysis.

4.4) Data Tamer is a very promising method for dealing with big databases, it includes getting rid of duplicates, cleaning the data, crowdsourcing, and methods in machine learning. The results are also very good compared to the existing methods, and many different fields would find Data Tamer very useful. Based on user’s response, it is also easy to use, with a grade of 2.6/3.