Second International Workshop on Preservation of Evolving Big Data - Panel on Big Data Quality

Angela Bonifati

University of Lyon 1
Liris – CNRS, France

March 15, 2016

Table of contents

image by Camelia.boban

Q_1: Risk assessment of poor data quality

Poor data entails poor data analysis

- In our ongoing research project (MedClean[a]), we work on medical data involving patient data, DNA sequencing data and medical images (as issued by microscopes), that may exhibit inconsistency, incompleteness and noise.
- Consequently, the diagnoses built on top of them might be affected.
- Very often, the data cannot be entirely cleaned due to many factors: the absence of authoritative sources, the inherent format of data, privacy issues.
- Can we characterize the uncertainty of such data and thus express a confidence measure of the subsequent analysis?

[a] Défi CNRS Mastodons 2016, MedClean (PI: AngelaBonifati)

Q_2:What are the main quality factors of Big Data?

Quality depends on the format

- Because of one of the Vs of Big Data, namely Variety, many diverse data formats need to co-exist, each of which requires a specific notion of quality.

Relational tuples, Graphs, Time series, Images $\cdots$

- if we restrict to relational tuples, we can identify measures/conceive techniques to ensure data quality (research is somehow mature);
- if we go beyond the relational models, the setting becomes less clear: for graphs, for instance, do we have a notion of quality? for images, is the quality related to its resolution and its precision? what about time series or DNS sequencing data?

Q_3: Challenges for Big Data Quality: the FOUR Vs

- Volume: scale factor of data involved in the cleaning processes drastically changes (e.g. images of the order of Terabytes are too voluminous to be manipulated by physicians/biologists, difficult to browse etc.)
- Velocity: time series from clinical data are an example of data in which velocity is rather critical (of the order of Petabytes...);
- Variety: already mentioned above; an important aspect of our project is about annotation and transformation across different formats.
- Veracity: trust is very important in health-care decisions and very difficult to guarantee on massive datasets. Can we isolate query-driven snippets on datasets for which we can provide guarantees of quality and trust?

Q_4: Quality Assessment and Decision Making (Part I)

- As mentioned before, sometimes we are unable to perform data cleaning at best for various reasons.
- Can we use (probabilistic?) approaches to measure the uncertainty of data so that we can also quantify the uncertainty (and the cost) of decision making processes?
- Even when data cleaning attains high values of precision and recall, we may need to quantify the uncertainty for instance because the involved data will be part of further processing and integration in the sequel and we want to keep track of its incompleteness.
- Finally, for privacy reasons, we may be in a situation in which we cannot clean the original data sources. We need to come up with assessment and decision making methods that work in a privacy-preserving manner.

Q_4: Can Data Quality be ignored? (Part II)

- Which threshold depends on the actual data formats and on the subsequent tasks of the lifecycle, as well as on the final analysis that needs to be performed.
- Maybe in some cases, it can but still we need to annotate data with suitable indicators.
- At what extent can data quality be ignored? We actually do not know...

Conclusion

Big Data Quality

- **Data quality**: design of quality-conscious methods.
- **Interactions between the Vs**: several open problems out there!

State-of-the-art and directions of research

- Existing large-scale data cleaning methods for relational databases, entity resolution for graphs....
- Combinations of data formats: are we ready?