Data Cleansing at the Data Entry to assert Semantic Consistency

Prof. Hans-J. Lenz,
Inst. of Information Science
Inst. of Statistics and Econometrics,
Freie Univ. Berlin, Germany,

We discuss edits or data validation rules applied to OLTP and OLAP data.

We start with a set of rather primitive edits, logical edits based on the well-known, but quite restrictive Boolean calculus, cf. Boskovits (2008), which is defined on binary symbolic data spaces, and probabilistic edits which are based on a measurable space (a set and a corresponding sigma algebra) allowing the application of the classical probability theory.

Next, we turn to arithmetic edits which consider numerical constraints as a non-linear equation system defined on a metric data space spanned by all given non-primary key attributes from data records captured at time of data entry. As an example think of definitions or balance equations like U = M * P with Sales U, Quantity M, price per unit P) or F= M * B (force F, mass M and acceleration B). It is assumed that each data item is crisp.

A much more 'natural' class of validation rules (or semantic integrity constraints) is defined on a multivariate metric space, where we can consider constraints and/or upper or lower bounds for specific attributes. Generally speaking, there exist at least two different approaches: Probability Theory mainly under a normal or Gaussian regime (Lenz and Rödel (1991)) or Possibility Theory (Fuzzy Set Theory), cf. Lenz and Müller (1998). There is empirical evidence that both approaches are practically equivalent iff all equations are linear and the cross-correlations between variables are ignorable. This is true because the embedded principles are the same: folding in Probability Theory (Johnson and Kotz (1972)) and the "Extension Principle" in Fuzzy Set Theory, Zadeh (1965). Note, however, that the interpretation of such operations may vary, cf. Dubois and Prade (2003).

Finally, as very restrictive assumptions are underlying both approaches, we can relax a Gaussian regime and linearity's of equations just by applying MCMC-techniques with special emphasis on the Particle Filter Theory (Metropolis-Hastings simulation algorithm), cf. Chip (2004), Köppen and Lenz (2005) and Köppen (2008). Evidently, despite of the sampling error, we can validate imprecise and uncertain data at the data entry of databases with respect to the semantic consistency in a methodologically sound and experienced way.

All algorithms are implemented by the "Business Intelligence Group" at Freie Universität of Berlin, Germany. The CASE tools used are JAVA Eclipse for the GUI, MS SQL server or IBM DB2 server, and the functional or matrix language R for number crunching.

****