

Information Quality



Ami Motro
Igor Rakov

Motivation

- Information products (like hardware products) need certified quality specifications.
- With quality specifications, we may
 - ▶ Estimate the **quality of answers** issued by an information source.
 - ▶ Select the **best answer** to a query, among several candidate answers obtained from multiple overlapping information sources.
 - ▶ Synthesize a **single answer** from several inconsistent candidate answers obtained from multiple overlapping information sources
- Information quality has been around informally in claims such as
 - ▶ "This information is accurate at time of distribution, and we reserve the right to change any information at any time thereafter."
 - ▶ "This mailing list is guaranteed to include at least 85% of those likely to buy a new car in the next 6 months."
 - ▶ "The information in this directory is estimated to be 93% correct."

Solution Methodology:

1. Quality Metrics

- Dual metrics for specifying information quality:
 - ▶ **Completeness**: measures to what degree the information includes **the whole truth**.
 - ▶ **Soundness**: measures to what degree the information includes **nothing but the truth**.
- All specifications are derived from statistical samples of the information products.

Solution Methodology:

2. Goodness Basis

- Specifications must reflect the variability of quality that is possible within the same information product:
 - ▶ The information concerning **color and pattern** is not as accurate as the information concerning **size or weight**.
 - ▶ The completeness of lists of **suppliers in the northeast** is much higher than the **overall completeness**.
- Developed algorithms for partitioning the information into areas of homogeneous quality:
 - ▶ Classification And Regression Trees (CART).
 - ▶ Homogeneity Indexes (Gini).
- Measured Goodness Basis: Vector of quality estimations for each homogeneous area:
 - ▶ Established by manual verification of the sample points.

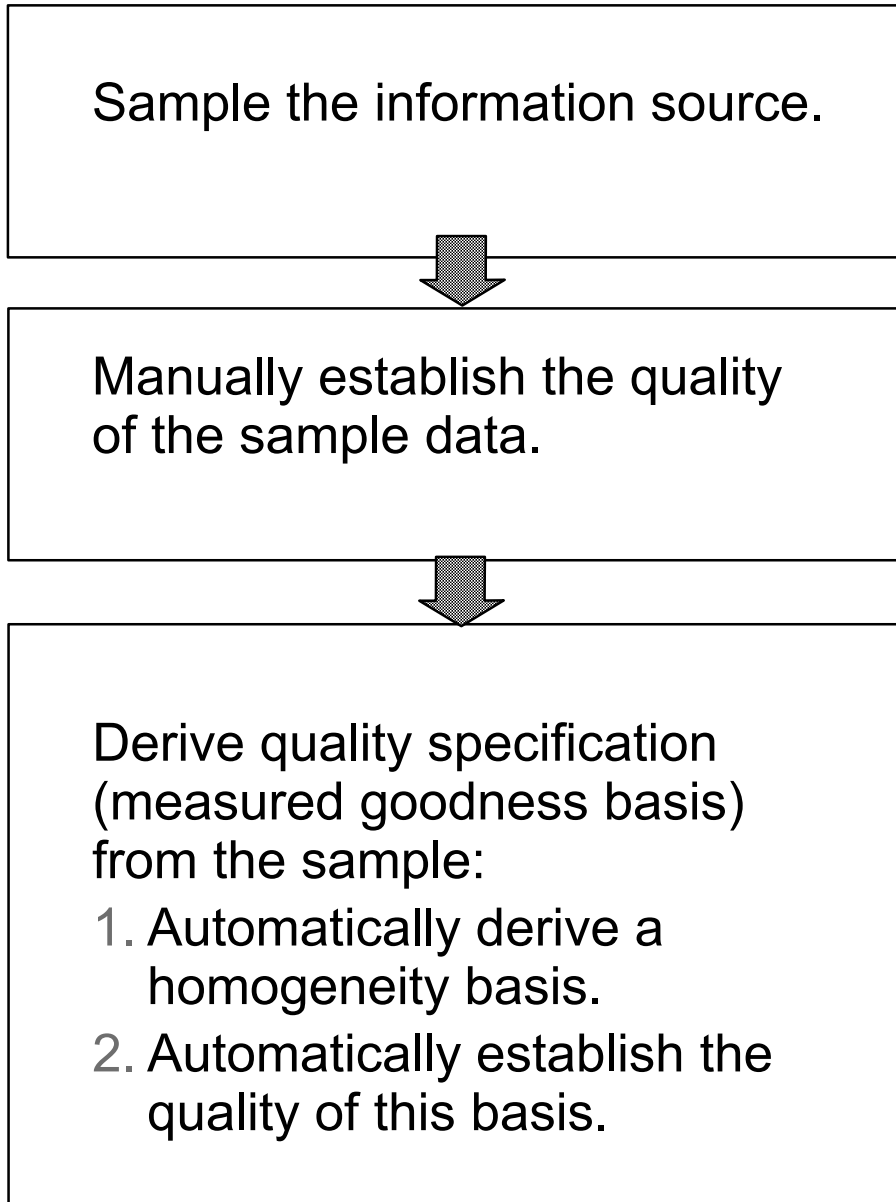
Solution Methodology:

3. Query Processing

- Extended query processing:
 - ▶ Processing of every retrieval operator is extended to deliver both an answer and a (measured) goodness basis for the answer.
 - ▶ Allows chaining of retrieval operators.
 - ▶ For the final answer in a complex query, the overall quality is derived from the measured goodness basis.

The Overall Process

Measure (once):



Apply (continuously):

