2. Inconsistency Resolution

Types of Inconsistency

- *Intensional inconsistencies* (semantic differences). Examples:
  1. Structural differences (telephone numbers stored in one field or in several fields).
  2. Unit differences (Dollars vs. Euros).
  3. Difference in semantics of attributes (yearly salary vs. monthly salary).
- *Extensional inconsistencies* (data conflicts):
  1. Surface only after all intensional inconsistencies have been resolved.
  2. Two different values for the date of birth of same person.
  3. Subject received much less attention.

Classification of Solutions to Extensional Inconsistencies

Extensional inconsistencies can be handled in different ways:

1. **Multi-answer**: The complete set of inconsistent answers (*disjunctive answer*). Raw information that should be resolved outside the database.
2. **Ranked answer**: The complete set of answers, but ranked according to likelihood of being correct. Usually, ranking derived from rate of recurrence.
3. **Random Answer**: Single value selected at random. Useful when differences among alternatives considered inconsequential.
4. **Preferred answer**: The top value in a ranked answer.
5. **Fused answer**: A new value synthesized from the set of answers. Normally, the fusion formula is provided by "experts" who know the sources.

Weaknesses of Current Approaches

1. Multi-answer/Random are *naive* solutions that require no further investigation. We focus on the Ranked/Preferred and Fusion solutions.

2. Ranked/Preferred: Because these solutions are based only on voting, they are essentially useless when the set of alternatives is small, or the degree of recurrence is low.

3. Fusion: There is no measure to indicate if the fusion is any *improves* on the original values.

4. Fusion: There is no proof that the expert prescribed the "best" fusion.

Our Approach to Inconsistency Resolution

Assumptions

1. Assume a set of *performance measures* that quantify the performance of the sources: accuracy, cost, etc.
2. Assume each alternative answer is associated with a value for each performance measure.
3. Assume a *utility function* that expresses overall value to individual users by means of a linear combination of the performance measures.

Expected Advantages

1. Ranking/Preferred: Define ranking based on utility.
2. Fusion: Calculate the utility of the fusion and check if it exceeds the utility of each of the original values.
3. Fusion: Find the optimal fusion (with highest utility).

Performance Measures

1. Recentness (t): The time in which the information was published. Basically, the timestamp.
2. Cost (c): The expense (download seconds, access fee) of materializing the answer.
3. Availability (v): The probability that the source will be available when needed.
4. Accuracy (v): Assume database values are estimates of the true value and have a normal distribution around the stored value; the standard deviation is the measure of accuracy.
5. Priority (p): A preference based on past performance, or a level of authority granted by a certifying agency.
6. Quality (q): Essentially, any specification which the data is warranted to meet or exceed.

Other possible: We argue here more for the *approach*, than for individual parameters.

Utility

Utility is a linear combination of the performance measures.

1. Assume performance measures $p_1, p_2, \ldots, p_m$.
2. Assume weights $w_1, w_2, \ldots, w_m$
   $0 \leq w_i \leq 1$, $\sum_{i=1}^{m} w_i = 1$.
3. Utility: $u = \sum_{i=1}^{m} w_i \cdot p_i$

Ranking

Straightforward:

1. Assume the inconsistent values $x_1, x_2, \ldots, x_n$.
2. The utilities $u(x_1), u(x_2), \ldots, u(x_n)$ are calculated.
3. Ranked Answer: The values are sorted according to their utility.
4. Preferred answer: The value with the highest utility.

Fusion

Fusion is a linear combination of the given values.

1. Assume (numerical!) values $x_1, x_2, \ldots, x_n$.
2. Assume coefficients $a_1, a_2, \ldots, a_n$
   $0 \leq a_i \leq 1$, $\sum_{i=1}^{n} a_i = 1$.
3. Fusion: $x = \sum_{i=1}^{n} a_i \cdot x_i$

Performance of the Fusion

To compute the utility of the fusion $x = \sum_{i=1}^{n} a_i \cdot x_i$ we must derive each of its performance measures.

1. Recentness $t(x) = now$
2. Cost* $c(x) = \sum_{i=1}^{k} \cdot c(x_i)$
3. Availability* $v(x) = \prod_{i=1}^{k} v(x_i)$
4. Accuracy $s(x) = \sqrt{\sum_{i=1}^{n} a_i^2 \cdot s^2(x_i)}$
5. Priority $p(x) = \sum_{1=1}^{n} a_i \cdot x_i$
6. Quality* $q(x) = \min_{i=1}^{k} q(x_i)$

---

* $a_1, \ldots a_k$ are assumed to be the positive coefficients.

Normalization of the Performance Measures

To facilitate finding the appropriate weights in the utility, each performance measure is normalized to be in the range $[0, 1]$, with 1 corresponding to "best" performance and 0 to "worst".

To make each measure a function of all n coefficients (whether zero or positive), we use:

1. Cost: $\lceil a_i \rceil \equiv$ "if $a_i = 0$ then 0 else 1".
2. Availability: $\max\{v_i(x), \lfloor (1 - a_i) \rfloor\} \equiv$ "if $a_i = 0$ then 1 else $v_i(x)$".
3. Quality: $\max\{q_i(x), \lfloor (1 - a_i) \rfloor\} \equiv$ "if $a_i = 0$ then 1 else $q_i(x)$".

Normalization of the Performance Measures (Cont.)

The normalized performance measures of the fusion:

1. recentness $t(x) = 1$
2. cost $c(x) = 1 - \sum_{i=1}^{n} \lceil a_i \rceil \cdot (1 - c(x_i))$
3. availability $v(x) = \prod_{i=1}^{n} \max\{v(x_i), \lfloor (1 - a_i) \rfloor\}$
4. accuracy $s(x) = 1 - \sqrt{\sum_{i=1}^{n} a_i^2 \cdot (1 - s^2(x_i))}$
5. priority $p(x) = \sum_{i=1}^{n} a_i \cdot p(x_i)$
6. quality $q(x) = \min_{i=1}^{n}\{\max\{q(x_i), \lfloor (1 - a_i) \rfloor\}\}$

Utility of the Fusion

$$u(x) = w_1 \cdot t(x) + w_2 \cdot c(x) + w_3 \cdot s(x) + w_4 \cdot p(x) + w_5 \cdot v(x) + w_6 \cdot q(x)$$

1. We regard fusion as an attempt to improve upon the initial values.
2. Hence, fusion is *justified* if $u(x) > \max_{i=1}^{n} u(x_i)$.
3. But even if the fusion is justified, it may not be the "best" fusion possible: The fusion formula prescribed by the "expert" may not be optimal (with respect to utility).

Optimizing the Fusion

The utility of the fusion $x = \sum_{i=1}^{n} a_i \cdot x_i$ is expressed as a function of the coefficients a_i:

$$
\begin{aligned}
u(a_1, a_2, \ldots, a_n) &= w_1 \cdot 1 \\
&+ w_2 \cdot (1 - \sum_{i=1}^{n} \lceil a_i \rceil \cdot (1 - c_i)) \\
&+ w_3 \cdot (1 - \sqrt{\sum_{i=1}^{n} a_i^2 \cdot (1 - s_i)^2}) \\
&+ w_4 \cdot \prod_{i=1}^{n} \max\{v_i, \lfloor (1 - a_i) \rfloor\} \\
&+ w_5 \cdot \sum_{i=1}^{n} a_i \cdot p_i \\
&+ w_6 \cdot \min_{i=1}^{n}\{\max\{q_i, \lfloor (1 - a_i) \rfloor\}\}
\end{aligned}
$$

There are methods and packages to optimize such functions.

Optimizing the Fusion: Example

Performance data on a multi-answer of 5 values:

Property (raw)	x_1	x_2	x_3	x_4	x_5
Recentness (timestamp)	10	20	30	30	60
Cost (cents)	80	50	30	10	140
Accuracy (standard deviation)	2.5	0.5	2	1	1.5
Availability (probability)	0.6	0.4	0.7	0.9	0.3
Priority (on a scale of 0–5)	4	2	5	1	3
Quality (on a scale of 0–10)	7	6	3	4	5

Property (normalized)	x_1	x_2	x_3	x_4	x_5
Recentness	0	0.053	0.105	0.105	0.263
Cost	0.258	0.161	0.097	0.032	0.452
Accuracy	0	0.800	0.200	0.600	0.400
Availability	0.6	0.4	0.7	0.9	0.3
Priority	0.8	0.4	1.0	0.2	0.6
Quality	0.7	0.6	0.3	0.4	0.5

Optimization Example (Cont.)

Five utility formulas and their corresponding optimum fusions:

Property	u_1	u_2	u_3	u_4	u_5
Recentness	0.167	0.250	0	0	0
Cost	0.167	0	0.333	0.500	1
Accuracy	0.167	0.250	0.333	0	0
Availability	0.167	0	0.333	0	0
Priority	0.167	0.250	0	0.500	0
Quality	0.167	0.250	0	0	0

No.	**Optimal fusion**
u_1	$0.482 \cdot x_2 + 0.303 \cdot x_3 + 0.215 \cdot x_5$
u_2	$0.148 \cdot x_1 + 0.293 \cdot x_2 + 0.337 \cdot x_3 + 0.222 \cdot x_4$
u_3	$0.735 \cdot x_2 + 0.184 \cdot x_4 + 0.082 \cdot x_5$
u_4	x_3
u_5	x_4