Utility-based Resolution of Data Inconsistencies

Amihai Motro   Philipp Anokhin   Aybar C. Acar

Proceedings of IQIS 2004, International Workshop on Information Quality in Information Systems, pages 35-43.

Outline

- Problem Definition
- Assumptions
- Types of Inconsistencies
- Integration Strategies
- The Utility Approach
- Utility-based Fusion
- Conclusions

The Problem

- Virtual Database Systems
  - Systems that integrate multiple, independent information sources
- Occasionally, two or more sources may contain different values that purport to represent the same real world value.
  - Therefore certain queries on virtual databases may prove to be inconclusive.

Inconsistencies

- **Intensional inconsistencies**
  - Structural differences (phone numbers or addresses stored in different formats)
  - Unit differences (Dollars vs. Euros)
  - Semantic differences (yearly vs. monthly salaries)
- **Extensional inconsistencies**
  - Surface only after all intensional inconsistencies have been resolved.
  - Two different values for the date of birth of the same person.
  - Subject has received much less attention.

Integration Strategies

- **Multi-answer:** The complete set of inconsistent answers. The inconsistency is resolved outside the database.
- **Ranked answer:** The complete set of answers, but ranked according to the likelihood of being correct. Usually the ranking is based on rate of recurrence.
- **Random answer:** Single value selected at random. Appropriate when differences between alternates are negligible or inconsequential.
- **Preferred answer:** The top value in a ranked answer.
- **Fused answer:** A synthetic answer created by combining alternates. Normally, the fusion method is decided by experts.

Weaknesses of Current Approaches

- **Multi-answer** and **random** are *naïve* solutions that require no further investigation.
- **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.
- **Fusion:**
  - There is no measure to indicate if the fusion improves on the original values.
  - There is no proof that the expert prescribed the *best* fusion.

The Utility Approach

- Assumptions
  - Assume a set of *performance measures*.
    - Analogous to quality dimensions.
    - Accuracy, cost, recentness etc.
  - Assume each answer is associated with a vector of performance measures.
  - Assume a *Utility Function* that expresses overall value to the data consumer by means of a linear combination of the measures.
- Expected Advantages
  - Ranking/Preferred: Perform ranking based on utility.
  - Fusion: Calculate the utility of the fusion and check if it exceeds the utility of the original values.
  - Fusion: Find the optimal fusion (maximum utility).

Performance Measures

- **Recentness** (*t*): The time at which the data was collected (i.e., timestamp) .
- **Cost** (*c*): The expense of materializing the answer (e.g., access fee, connection time etc.).
- **Availability** (*v*): The probability that the source will be available when needed.
- **Accuracy** (*a*): Assuming a normal distribution of possible data values around the true value, accuracy is the standard deviation.
- **Priority** (*p*): A measure of preference dictated by past performance or expert certification.
- **Quality** (*q*): A specification which the source guarantees that its data will meet or exceed.

Performance Measures

- The measures used here are examples, intended only serve to illustrate the general approach.
- Measures have certain properties:
  - Automatically determinable measures are preferred.
    - e.g., recentness, availability, priority.
  - In case human expertise is required, the measure is best determined at the level of source or at most relation, as opposed to individual data elements.

Utility

Utility is a linear combination of performance measures:

- Given performance measures $p_1, p_2 \ldots p_m$
  and weights $w_1, w_2 \ldots w_m$
  such that $0 \leq w_i \leq 1$ and $\Sigma^m{}_{i=1}\, w_i = 1$

Utility: $u = \Sigma^m{}_{i=1}\, w_i\, p_i$

Ranking

- Assume inconsistent data values $x_1, x_2, x_3 \ldots x_n$ from different sources.
- Calculate utilities $u(x_1), u(x_2), u(x_3) \ldots u(x_n)$.
- Ranked Answer:
  - The values are sorted and returned according to descending utility.
- Preferred Answer:
  - The value with the highest utility is returned as the preferred answer

Fusion

- Assume inconsistent data values $x_1, x_2, x_3 \ldots x_n$ from different sources.
- Assume fusion coefficients $a_1, a_2, a_3 \ldots a_n$ such that:
  - $0 \leq a_i \leq 1$
  - $\Sigma^{n}_{i=1} a_i = 1$

Fusion: $x_f = \Sigma^{n}_{i=1} a_i x_i$

Performance Measures of Fused Values

To compute the utility of the fused value we need fused performance measures:

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

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

Normalization of Performance Measures

To facilitate the optimization of the fusion the performance measures are normalized to between 0 (worst) and 1 (best):

If $a_i = 0$ then 0 else 1

If $a_i = 0$ then 0 else $v_i(x)$

recentness $t(x) = 1$

cost $c(x) = 1 - \sum_{i=1}^{n} \lceil a_i \rceil \cdot (1 - c(x_i))$

availability $v(x) = \prod_{i=1}^{n} \max\{v(x_i), \lfloor (1 - a_i) \rfloor\}$

accuracy $s(x) = 1 - \sqrt{\sum_{i=1}^{n} a_i^2 \cdot (1 - s^2(x_i))}$

priority $p(x) = \sum_{i=1}^{n} a_i \cdot p(x_i)$

If $a_i = 0$ then 0 else $q_i(x)$

quality $q(x) = \min_{i=1}^{n}\{\max\{q(x_i), \lfloor (1 - a_i) \rfloor\}\}$

Utility of the Fusion

- **Definition**:
  $u(x_f) = w_1 \cdot t(x_f) + w_2 \cdot c(x_f) + w_3 \cdot s(x_f) + w_4 \cdot p(x_f) + w_5 \cdot v(x) + w_6 \cdot q(x_f)$
  - We regard fusion as an attempt to improve upon the initial values.
  - Hence, fusion is justified if $u(x) > max^n_{i=1} u(x_i)$.
  - However, even if the fusion is justified, it is not necessarily the best option.
    - The expert defined fusion formula may not be optimal with respect to utility.
  - Hence the utility of the fusion is a linear function
    - We can optimize the mixing coefficients (a_i) to get the highest utility.
    - Possible to solve efficiently using linear programming algorithms (e.g., simplex).

Fusion Optimization Example

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

↓

Measure (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	0.8	0.2	0.6	0.4
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

- The following optimizations are possible for this set of answers:

Fusion	Formula
u_1	$0.483 \cdot x_2 + 0.303 \cdot x_3 + 0.215 \cdot x_5$
u_2	$0.148 \cdot x_1 + 0.293 \cdot x_2 + 0.337 \cdot x_3 + 0.222 \cdot x_1$
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

- With these performance values:

Measure	u_1	u_2	u_3	u_4	u_5
Recentness	0.167	0.250	0	0	0
Cost	0.167	0	0.333	0.5	1
Accuracy	0.167	0.250	0.333	0	0
Availability	0.167	0	0.333	0	0
Priority	0.167	0.250	0	0.5	0
Quality	0.167	0.250	0	0	0

Conclusion

- A method to help solve data inconsistencies in information coming from multiple sources.
  - A model that associates metadata with each source.
  - Six sample measures considered, others might be possible.
  - These measures are combined into a linear objective function: *Utility*.
- The *true* value might not be one of the answers obtained from the sources.
  - It is possible that the true value is a “mixture” of the answers: *Fusion*.
  - The utility of such a fusion may exceed that of individual components.
  - It is also possible to find an optimal fusion using the utility of such fusion as the objective function.
- Future Directions
  - Consider non-numeric values and their fusion:
    - Parsing non-numeric values into sub-components and recombining an answer using recurrence as a guide.