Evaluating and Aggregating Data Believability across Quality Sub-Dimensions and Data Lineage

Nicolas Prat
Stuart E. Madnick

Working Paper CISL# 2007-09

December 2007

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Evaluating and Aggregating Data Believability across Quality Sub-Dimensions and Data Lineage

Nicolas PRAT
ESSEC Business School
Avenue Bernard Hirsch
BP 50105 - 95021 Cergy Cedex - FRANCE
prat@essec.fr

Stuart MADNICK
MIT Sloan School of Management
30 Wadsworth Street – Room E53-321
Cambridge MA 02142 - USA
smadnick@mit.edu

Abstract

Data quality is crucial for operational efficiency and sound decision making. This paper focuses on believability, a major aspect of data quality. The issue of believability is particularly relevant in the context of Web 2.0, where mashups facilitate the combination of data from different sources. Our approach for assessing data believability is based on provenance and lineage, i.e. the origin and subsequent processing history of data. We present the main concepts of our model for representing and storing data provenance, and an ontology of the sub-dimensions of data believability. We then use aggregation operators to compute believability across the sub-dimensions of data believability and the provenance of data. We illustrate our approach with a scenario based on Internet data. Our contribution lies in three main design artifacts (1) the provenance model (2) the ontology of believability sub-dimensions and (3) the method for computing and aggregating data believability. To our knowledge, this is the first work to operationalize provenance-based assessment of data believability.

1. Introduction

Data quality is crucial for operational efficiency and sound decision making. This paper focuses on the assessment of data believability. Wang and Strong [1] define this concept as "the extent to which data are accepted or regarded as true, real and credible". Their survey shows that data consumers consider believability as an especially important aspect of data quality. Furthermore, the issue of data believability is particularly relevant in the context of Web 2.0 with the advent of mashups, defined as "web applications that combine data from more than one source into an integrated experience" (source: http://www.wikipedia.org). Data users need to be able to assess the believability of the data sources and of the data resulting from their combination.

Clearly, the believability of a data value depends on its origin (sources) and subsequent processing history. In other words, it depends on the data provenance (aka lineage), defined in [2] as "information that helps determine the derivation history of a data product, starting from its original sources". Data provenance is the subject of several papers (see [2] for a literature review on this topic). Several papers, e.g. [3], stress the relationship between data quality assessment and provenance information. However, a computational model of provenance-based data quality, and more specifically believability, is still missing. Our work aims at addressing this issue.

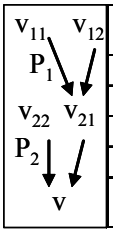
2. Data provenance and lineage

We present the main concepts of our provenance model, defined to represent and store provenance information for subsequent computation of believability. The complete UML [4] model is described in [5], the key elements are explained in this section. Data values are central to our model. Our research focuses on numeric, atomic data values (complex values, e.g. tuples or relations, will be considered in further research). A data value may be a source or resulting data value, where a resulting data value is the output of a process run (instantiation of a process). Processes may have several inputs and only have one output. Processes are executed by agents. This concept also represents the providers of the source data values. Our model uses the database concepts of valid time and transaction time (more details are provided in [5]). Finally, the concept of trustworthiness is essential for assessing data believability. It is defined for an agent, for a specific knowledge domain. It is measured by a trustworthiness value. The

determination of these values is outside the scope of our work. We assume they are computed or obtained from outside sources, e.g. reputation systems.

The terms "data lineage" and "provenance" are often used interchangeably. We define the lineage of a data value v (noted Lineage(v)) as a labeled, directed acyclic graph representing the successive data values and processes leading to v. The data values are the vertices of the graph (noted V(Lineage(v))), and the processes are the labeled edges. Therefore, lineage is a specific, graph-based view on provenance information, focused on the chaining between data values and processes.

To illustrate our representation of data lineage and provenance, we will use the following scenario: A communication group considers launching a new TV channel in Malaysia and Singapore, aimed at the Indian community. The group needs to know the total Indian population in Malaysia and Singapore. This figure is computed from source data found on Internet. We wish to assess the believability of this figure (the value v). The lineage of v is represented in the left part of Figure 1. In this case, P_1 is the multiplication and P_2 the sum; the values in v's lineage and their characteristics (provenance information) are shown in the right part of Figure 1.



Id	Data	Value	Transaction time	Start valid time	End valid time	Provided By
v_{11}	Total population of Malaysia in 2004	25 580 000	31-Dec-05	1-Jan-04	31-Dec-04	Malaysian Dpt of Stats
v_{12}	% Indian population in Malaysia in 2004	7.1	31-Dec-04	1-Jan-04	31-Dec-04	CIA
v_{21}	Indian population in Malaysia in 2004	1 816 180	12-Feb-06	1-Jan-04	31-Dec-04	$P_1(v_{11},v_{12}) = v_{11} * v_{12}$
v_{22}	Indian population in Singapore in 2006	319 100	30-Jun-06	1-Jan-06	31-Dec-06	Singapore Dpt of Stats
v	Indian population in Malaysia and Singapore in 2006	2 135 280	1-Jun-07	1-Jan-06	31-Dec-06	$P_2(v_{21},v_{22}) = v_{21} + v_{22}$

Figure 1. Example scenario

3. Ontology of believability sub-dimensions

A dimension of data quality, believability is itself decomposed. [6] proposes three sub-dimensions, namely believability: (1) of source, (2) compared to internal commonsense standard, and (3) based on temporality of data. Our ontology of believability refines this typology, decomposing the three initial sub-dimensions of believability :

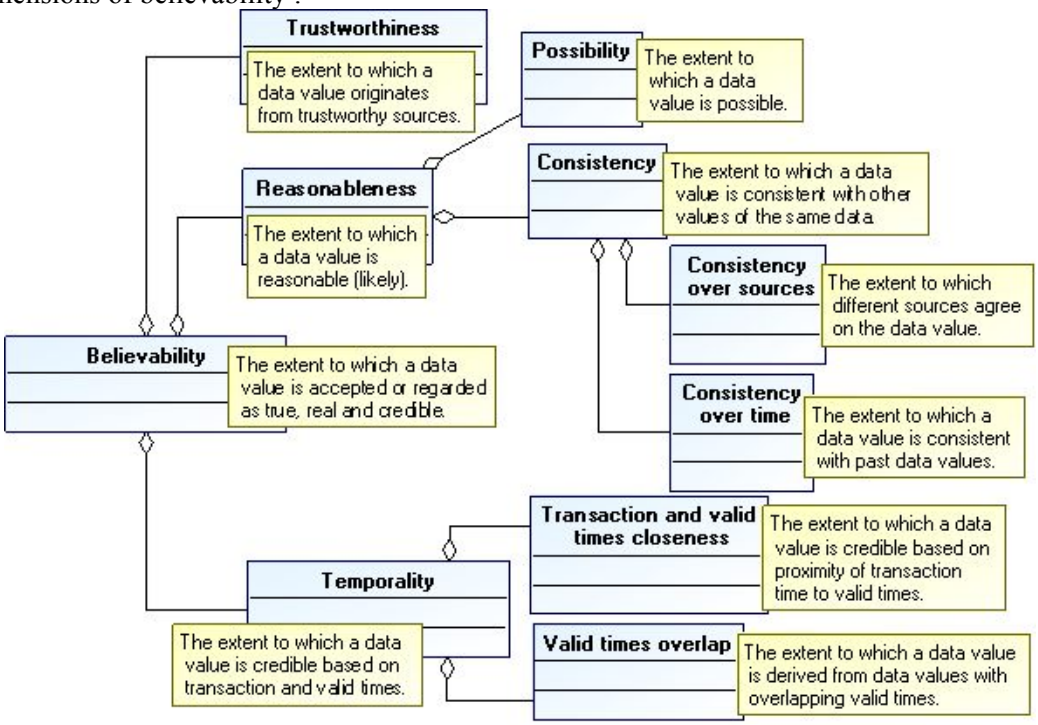


Figure 2. Sub-dimensions of believability (UML notation)

In [5], we have defined specific metrics for each of the 6 elementary sub-dimensions of data believability. Each of these metrics is based on the semantics of the believability sub-dimension for which it is defined. The present paper focuses on aggregation and, to this end, applies the theory of aggregation operators.

4. Aggregation operators

The formal theory of aggregation operators [7] [8] consolidates contributions from various fields, primarily fuzzy logic. We summarize the main characteristics of these operators.

An aggregation operator with n arguments is a function $A_{(n)}$: $[0,1]^n \rightarrow [0,1]$ verifying the following conditions:

1. *Monotonicity*.
2. *Boundary conditions*, namely (2.1) $A_{(n)}(0,\ldots,0)=0$, (2.2) $A_{(n)}(1,\ldots,1)=1$, and (2.3) $\forall\ x \in [0,1]$, $A_{(1)}(x)=x$.

Among the properties that aggregation operators may or may not possess, we should mention the following characteristics, which are relevant for aggregating data believability:

3. *Continuity*, which is an important property when combining (quality) measures. Continuity ensures that small errors in input values cannot cause a big error in the aggregated result. To this respect, it should be noted that the arithmetic means is the most stable aggregation operator under possible input errors.
4. *Neutral element* and *annihilator* (aka absorbing element). For example, for the operator $\prod$ (product), 1 is the neutral element while 0 is the annihilator.
5. *Dual*: the dual of $A_{(n)}$, noted $A_{(n)}d$, is defined by: $A_{(n)}d(x_1,\ldots,x_n)=1 - A_{(n)}(1-x_1,\ldots,1-x_n)$.
6. *Idempotency*: an aggregation operator is idempotent if $A_{(n)}(x,\ldots,x)=x$. Min and Max are examples of idempotent operators, while $\prod$ is a counter-example.
7. *Reinforcement*: an operator exhibits downward reinforcement if when all the x_i are disaffirmative (i.e. inferior to the neutral element), $A_{(n)}(x_1,\ldots,x_n) \leq \mathrm{Min}_i\ (x_i)$. $\prod$ is an example of downward reinforcement operator. Upward reinforcement operators are defined similarly. Uninorms, defined by Yager and Rybalov, may exhibit both upward and downward reinforcement. The 3-$\prod$-operator (applied below) is an example.
8. *Different weights of criteria*: If the x_i have different weights, and if these weights are known or can be determined, an aggregator such as the weighted arithmetic means may be used. OWA operators (Ordered Weighted Average), introduced by Yager, are a variation of the weighted arithmetic means: the weights are not pre-assigned to criteria, but affected dynamically based on the value of criteria: the x_i which has the highest value gets the highest weight, etc.

Unfortunately, several of these characteristics are antagonistic. For example, idempotency is not compatible with strict reinforcement. Consequently, the choice of the adequate operator should be based on the required properties, depending on context. In the following section, we make some initial suggestions regarding choice of aggregation operators to be used.

5. Aggregating data believability across quality sub-dimensions and data lineage

The believability of a data value v (noted B(v)) is assessed by computing the "elementary" measures of believability and then aggregating the results across the sub-dimensions of believability and the lineage of v. An "elementary" measure of believability is the computation of an elementary sub-dimension of data believability for a data value in the lineage of v. Therefore, we have:

$$B(v) = \mathrm{Aggregation}_{Q_j \in \text{ elementary sub-dimensions of data believability}} (\mathrm{Aggregation}_{v_i \in \text{ v's Lineage}} (Q_j(v_i)))$$

The measures for computing the $Q_j(v_i)$ are developed in [5]. It should be noted that only one operator is required to aggregate believability across the lineage of v, while several different operators may be used to perform aggregation across the different sub-dimensions of believability (depending on the level in the ontology of Figure 2). Observe that aggregation is performed first across the lineage of v and then across the sub-dimensions of believability. The reverse is not possible, since the weights used for

aggregating believability across data lineage may differ depending on the considered believability sub-dimension (as explained below).

For aggregating believability across the graph of v's lineage, two types of weights emerge:

- For a process P_k in v's lineage, the weight of an input data value v_{ki} on the output value is computed similarly to [9]: we compute the derivative dP_k/dx_{ki} (the "slope") and multiply by v_{ki}. (We assume that the partial derivatives of P_k exist, and take absolute values to avoid negative weights). The weights need to be composed along the whole process chain.
- Furthermore, as often proposed in graph-based algorithms, we introduce an "attenuation factor" [10], reflecting the intuition that the influence of a vertex on another decreases with the length of the path separating the two vertices. In our case, the influence of v_{ki} on the believability of data value v diminishes with the length of the path from v_{ki} to v. Note that contrary to the first type of weight, the attenuation factor may differ depending on the sub-dimension of believability, reflecting different degrees of attenuation depending on the semantics of the sub-dimension of believability and its mode of computation.

Combining these two types of weights, we compute an influence vector $G_j(v)$, reflecting the influence of v's lineage on the aggregated believability of v (since the influence vector may depend on the sub-dimension of believability Q_j, it is suffixed by j):

From the graph Lineage(v), build the matrix M(v) defined as follows:
*M(v) is a square matrix of order Card(V(Lineage(v))) * Card(V(Lineage(v)))*
The rows/columns of M(v) represent the vertices of Lineage(v). (The last row/column represents v.)
Call v_r *and* v_c *the vertices corresponding to row r and column c respectively.*
The content of element $M_{r,c}(v)$*, where r and c* $\in$ *1..Card(V(Lineage(v))), is defined as follows:*

If $\exists$ *an edge e from* v_r *to* v_c *in Lineage(v)*
Then Call *P the label of e*.

$$M_{r,c}(v)= \left| \frac{dP}{dx_r}(v_r) * v_r \right|$$

Else $M_{r,c}(v)=0$
Endif
For each *column c, divide each element* $M_{r,c}(v)$ *by the sum of the elements of column c.*

Let $N_j(v)= \sum_{k=0}^{length(Lineage(v))} (\gamma_j * M(v))^k$

($\gamma_j \in [0..1[$ *is an attenuation factor; length(Lineage(v)) is the length of the longest process chain from a source data value to v; the first term of the sum* $\sum$ *, i.e. for k=0, is the identity matrix).*

The final vector $G_j(v)$ *is the transpose of the last column of* $N_j(v)$*, divided by the sum of elements of this column in order to normalize weights. (U is the unit column vector).*
$O_j(v)= (N_j(v)\ [1..\ Card(V(Lineage(v)));\ Card(V(Lineage(v)))])^T$
$G_j(v)=(1/\ (O_j(v)\ *U))*\ O_j(v)$

Observe that in the algorithm above, the propagation of weights is performed through matrix multiplication, reflecting a known property of adjacency matrices [10]: the square of an adjacency matrix propagates links one level (the "friends of friends"), etc. We also note, following [10], that the attenuation factor γ_j should be such that $\gamma_j < (1/\lambda)$, where λ is the largest absolute value of any eigenvalue of matrix M(v).

Figure 3 illustrates the computation of the influence vector for our application scenario. The matrix M(v) is shown in the top, and the influence vector $G_j(v)$ for a particular sub-dimension of believability is shown in the bottom. Observe that M(v) has only 0s on the diagonal (reflecting that the same data value cannot be both the input and output of the same process run); also, M(v) is a triangular matrix, which may not necessarily be the case in general. The computed values for the elements of the matrices are shown,

assuming a value of 0.5 for γ_j (an arrow points from each computed value to the corresponding matrix element).

As an illustration of the computations of the values in Figure 3, consider the value of $M_{3,5}(v)$ (i.e. the value noted as c in Figure 3). From Figure 1, we get the values for v_{21} and v_{22}, and we know that $P_2(v_{21},v_{22})=v_{21}+v_{22}$. Therefore, the derivative $\frac{dP_2}{dx_{21}}(v_{21})=1$ and, similarly, $\frac{dP_2}{dx_{22}}(v_{22})=1$. It follows that $M_{3,5}(v)= (1*1816180) / (1*1816180 + 1*319100) = 0.85$

Note that in our example, M(v) being a triangular matrix with 0s on the diagonal, the eigenvalues are all 0 and do not provide an upper bound for the value of γ_j.

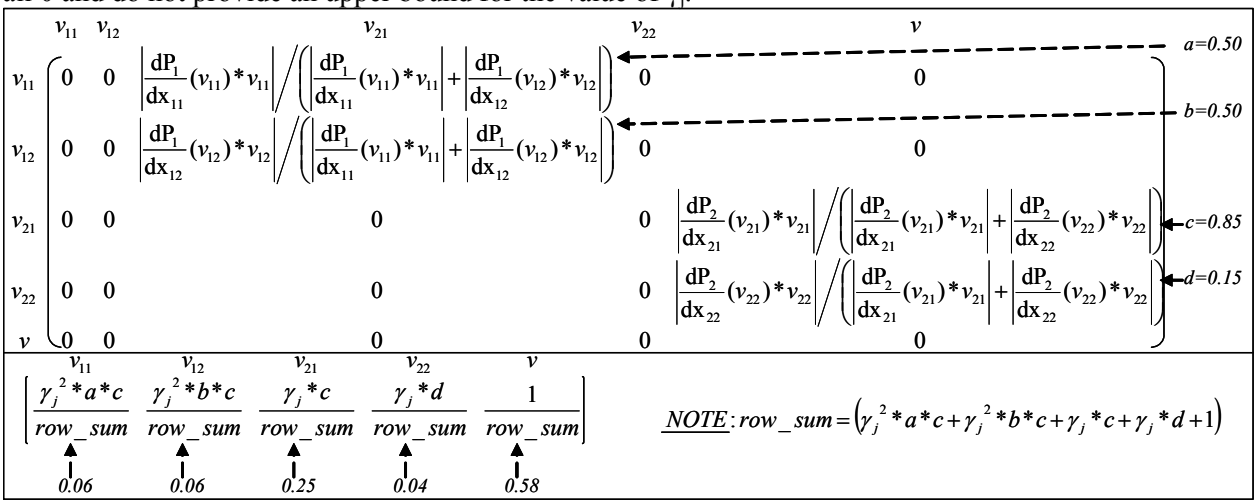


Figure 3. Application scenario: matrix M(v) (top) and example influence vector (bottom)

For a given elementary sub-dimension Q_j of data believability, we may now perform aggregation across the lineage of v. Since we have weights (computed in the influence vectors $G_j(v)$ above), the choice of the weighted arithmetic means seem natural. More precisely, we choose the dual of the weighted arithmetic means, following similar lines as [9], which combines errors based on derivatives. Errors are interpreted as lack of quality, hence the dual:

Call $MatQ_j(v)$ the column vector containing the $Q_j(v_i)$, where $v_i \in$ v's lineage (each row of $MatQ_j(v)$ represents a vertice v_i of Lineage(v), in the same order as in matrix M(v) above).

Aggregation$_{v_i \in \text{ v's Lineage}}(Q_j(v_i))=1-G_j(v)*(U - MatQ_j(v))$

We pursue with aggregation across the sub-dimensions of believability. To this end, as a general rule, we use the arithmetic means. This is justified by the fact that (as mentioned above) this operator is the most stable aggregation operator. Furthermore, we don't have weights for the sub-dimensions of believability (although in practice, the determination of these weights could be considered, e.g. by interviewing data users). We make two exceptions to the use of the arithmetic means: the first exception concerns aggregation along the reasonableness sub-dimension. Here, we use the OWA operator, counting the lowest value (among possibility and consistency) twice as much as the highest value. This choice reflects the particular semantics of possibility and consistency. We want to combine both possibility and consistency when computing reasonableness, but are more confident in small values. The second exception is the choice of the 3-$\prod$-operator at the highest aggregation level (final computation of B(v)). We make this choice because it provides a downward reinforcement effect for values under 0.5 (penalizing bad quality) and, conversely, an upward reinforcement effect for values above 0.5. We use the following notation convention for the sub-dimensions of believability: the index of each sub-dimension corresponds to its position in the hierarchy of sub-dimensions, as represented in Figure 2 (e.g. Q_1 stands for trustworthiness and Q_{222} stands for consistency over time):

$Q_1(v)=1-G_1(v)*(U – MatQ_1(v))$

$Q_{21}(v)=1-G_{21}(v)*(U - MatQ_{21}(v))$
$Q_{22}(v)=0.5*[(1-G_{221}(v)*(U - MatQ_{221}(v)))+ (1-G_{222}(v)*(U - MatQ_{222}(v)))]$
$Q_2(v)= 0.67*Min(Q_{21}(v), Q_{22}(v)) + 0.33*Max(Q_{21}(v), Q_{22}(v))$
$Q_3(v)=0.5*[(1-G_{31}(v)*(U - MatQ_{31}(v)))+ (1-G_{32}(v)*(U - MatQ_{32}(v)))]$
$B(v)= (Q_1(v)* Q_2(v)* Q_3(v)) / [Q_1(v)* Q_2(v)* Q_3(v) + (1-Q_1(v))* (1-Q_2(v))* (1-Q_3(v))]$

Table 1 applies our method to the example scenario. The table shows the elementary sub-dimensions of believability, with the corresponding values computed with the metrics of [5] (for the sake of brevity, the sub-dimension reasonableness and its sub-dimensions are ignored). For each of these sub-dimensions, a different value may be chosen for the attenuation factor γ, depending on the semantics of the sub-dimension. Here, we take the default value 0.5 for the attenuation factor, except for the sub-dimension Q_1 (Q_1 denotes trustworthiness, and the metrics defined in [5] already take full consideration of a value's lineage when computing its trustworthiness, making it unnecessary to consider this lineage again in the aggregation phase). Observe that for Q_{31}, aggregation across v's lineage decreases v's score, but not radically (due to the low weight of v_{22}). On the contrary, for Q_{32}, taking v's lineage into consideration improves v's score (although not radically).

Id	Q_1 (γ=0)	Q_{21}	Q_{221}	Q_{222}	Q_{31} (γ=0.5)	Q_{32} (γ=0.5)
v_{11}	0.9				1	1
v_{12}	0.8				1	1
v_{21}	0.85				1	1
v_{22}	0.9				0.159	1
v	0.857				1	0
Aggregation accross v's lineage	0.857				0.96	0.41

Final result : B(v)=0.929

Table 1. Application to the example scenario

6. Conclusion

Data quality is a crucial issue. A dimension of data quality, believability is itself an increasingly important topic. Although believability is closely related to data provenance and lineage, operational models and methods for evaluating and aggregating provenance-based believability are still missing. Our work is a contribution in this domain. Our computational, quantitative approach for assessing data believability is a step towards automation of this assessment, thereby contributing to the repeatability of the evaluation process and consistency of results. Future work will include further testing of the approach and of its scalability on (real world) case studies, and explore alternative aggregation operators.

7. References

[1] R. Wang and D. Strong, "Beyond Accuracy: what Data Quality Means to Data Consumers", Journal of Management Information Systems, vol. 12, no. 4, Spring 1996, pp. 5-34.

[2] Y.L. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science", SIGMOD Record, vol. 34, no. 3, September 2005, pp. 31-36.

[3] R. Wang, H. Kon, and S. Madnick, "Data Quality Requirements Analysis and Modeling", Proceedings of ICDE 1993, Vienna, Austria, April 1993.

[4] OMG, UML specification, v2.1.1, http://www.omg.org/technology/documents/formal/uml.htm

[5] N. Prat and S. Madnick, "Measuring Data Believability: a Provenance Approach", Proceedings of HICSS-41, Big Island, HI, USA, January 2008.

[6] Y. Lee, L. Pipino, J. Funk, and R. Wang, Journey to Data Quality, MIT Press, Cambridge, MA, 2006.

[7] D. Dubois, H. Prade, and R. Yager, "Merging Fuzzy Information", in J.C. Bezdek, D. Dubois, and H. Prade (eds), Fuzzy Sets in Approximate Reasoning and Information Systems, Kluwer, Boston, 1999.

[8] T. Calvo, A. Kolesarova, M. Komornikova, and R. Mesiar, "Aggregation Operators: Properties, Classes and Construction Methods", in T. Calvo, G. Mayor, and R. Mesiar (eds), Aggregation Operators – New Trends and Applications, Physica-Verlag, 2002.

[9] D.P. Ballou and H.L. Pazer, "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems", Management Science, vol. 31, no. 2, February 1985, pp. 150-162.

[10] L. Katz, "A New Status Index Derived from Sociometric Analysis", Psychometrika, vol. 18, no. 1, March 1953, pp. 39-43.