A Knowledge-based Approach to Assisting in Data Quality Judgement

Y. Jang

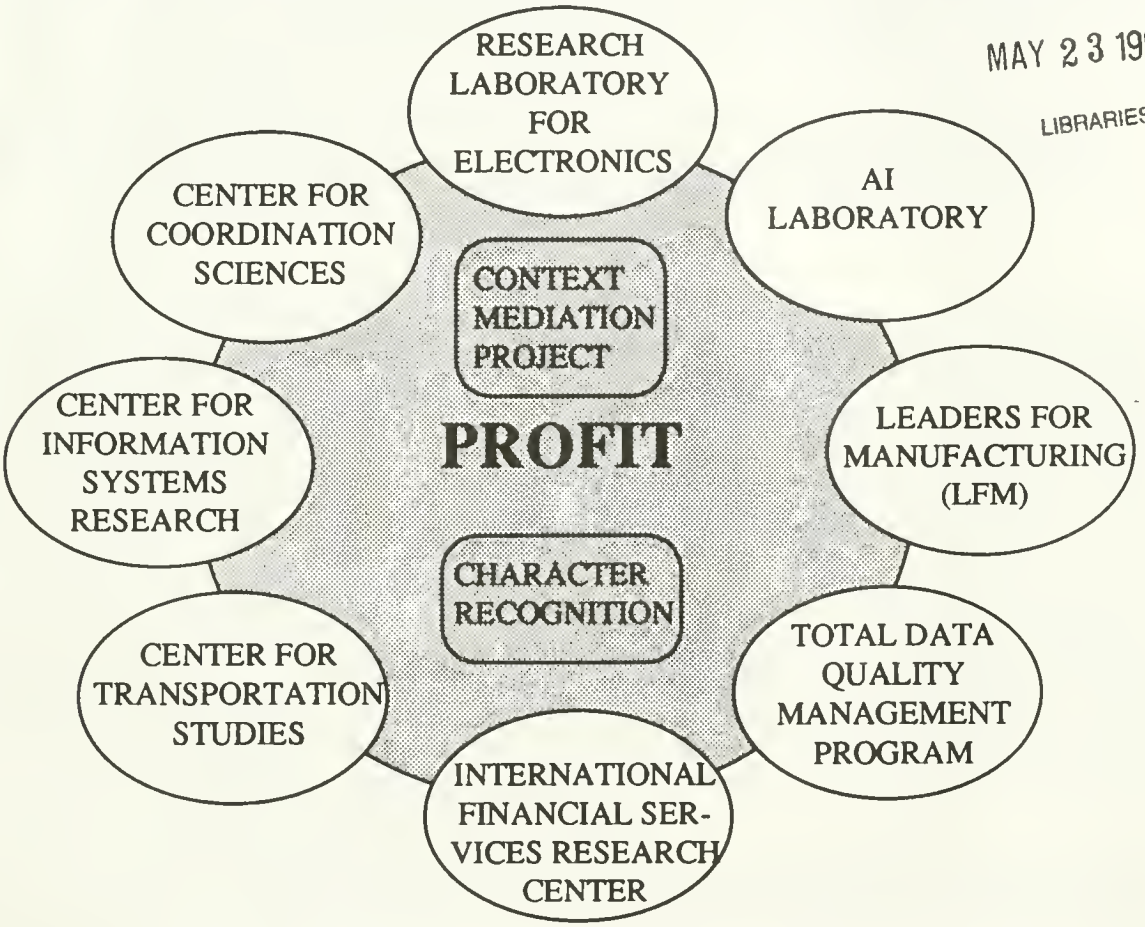
WP #3763 December 1992
PROFIT#92-02

Productivity From Information Technology
"PROFIT" Research Initiative
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
(617)253-8584
Fax: (617)258-7579

 The research described herein has been supported (in whole or in part) by the Productivity From Information Technology (PROFIT) Research Initiative at MIT.

Productivity From Information Technology (PROFIT)

The Productivity From Information Technology (PROFIT) Initiative was established on October 23, 1992 by MIT President Charles Vest and Provost Mark Wrighton "to study the use of information technology in both the private and public sectors and to enhance productivity in areas ranging from finance to transportation, and from manufacturing to telecommunications." At the time of its inception, PROFIT took over the Composite Information Systems Laboratory and Handwritten Character Recognition Laboratory. These two laboratories are now involved in research related to context mediation and imaging respectively.



In addition, PROFIT has undertaken joint efforts with a number of research centers, laboratories, and programs at MIT, and the results of these efforts are documented in Discussion Papers published by PROFIT and/or the collaborating MIT entity.

Correspondence can be addressed to:

The "PROFIT" Initiative
Room E53-310, MIT
50 Memorial Drive
Cambridge, MA 02142-1247
Tel: (617) 253-8584
Fax: (617) 258-7579
E-Mail: profit@mit.edu

A Knowledge-Based Approach to Assisting In Data Quality Judgment

(Extended Abstract)

Yeona Jang
Laboratory for Computer Science
Massachusetts Institute of Technology
Yeona@lcs.mit edu

Henry B. Kon
Sloan School of Management
Massachusetts Institute of Technology
hkon@mit.edu

Richard Y. Wang
Sloan School of Management
Massachusetts Institute of Technology
rwang@mit edu

Abstract

As the integration of information systems enables greater accessibility to data from multiple sources, the issue of data quality becomes increasingly important. This paper attempts to formally address the data quality judgment problem with a knowledge-based approach. Our analysis has identified several related theoretical and practical issues. For example, data quality is determined by several factors, referred to as *quality parameters*. Quality parameters are often not independent of each other, raising the issue of how to represent relationships among quality parameters and reason with such relationships to draw insightful knowledge about the overall quality of data.

In particular, this paper presents a *data quality reasoner*. The data quality reasoner is a data quality judgment model based on the notion of a "census of needs." It provides a framework for deriving an overall data quality value from local relationships among quality parameters. The data quality reasoner will assist data consumers in judging data quality. This is particularly important when a large amount of data involved in decision-making come from different, unfamiliar sources.

1. Introduction

As the integration of information systems has enabled data consumers to gain access to both familiar and unfamiliar data, there has been growing interest and activity in the area of data quality. Even if each individual data supplier were to guarantee the integrity and consistency of data, data from different suppliers may still be of different quality levels — due, for example, to different data maintenance policies. Unfortunately, as demonstrated in studies presented in the literature such as [Bonoma, 1985;Burnham, 1985;Johnson, 1990;Laudon, 1986], decisions made based on inaccurate or out-of-date data can result in serious economic and social damage. The problem of data quality is thus increasingly critical.

A majority of previous research efforts on data quality has focused on providing to data consumers "meta-data," *i.e.*, data about data, that can facilitate the judgment of data quality; for example, data source, creation time, and collection method. We refer to these characteristics of the data manufacturing process as *quality indicators* (see Table 1 for examples of quality indicators). Data-quality judgment is still, however, left to the data consumers. Unfortunately, information overload makes it difficult to analyze such data and draw useful conclusions about data quality. This paper seeks to assist data consumers in judging if the quality of data meets his or her requirements, by reasoning about information critical to data quality judgment.

Regarding data quality, this paper focuses especially on the problem of assessing levels of data quality, *i.e.*, the degree to which data meets desired characteristics of the data from the user's perspective. In considering the data quality assessment problem, our analysis has identified several theoretical and practical issues:

1) What are data quality requirements?
2) How can relationships between dimensions of these requirements be represented?
3) What can be known about overall data quality from such relationships, and how?

The study conducted on major US firms, in [Wang & Guarrascio, 1991], identified a relatively exhaustive list of requirements, such as timeliness and credibility. Such requirements are referred to as *quality parameters* in this paper (see Table 2 for examples of data quality parameters). Unfortunately,

Research presented in this paper was supported in part by the International Financial Services Research Center at MIT, in part by the National Heart, Lung, and Blood Institute under the grant number R01 HL33041, and in part by the National Institute of Health under the grant number R01 LM04493 from the National Library of Medicine.

requirements of data depend to largely on the intended usage of the data. For example, consider patient records. Availability of the records may be more important than accuracy to hospital administration, while to physicians accuracy is as important as availability of the records for effective patient management. The issue, then, is how to deal with such user- or application-specificity of quality-parameter relationships. This paper attempts to address this issue with a knowledge-based approach. This raises the issue of how to represent relationships among quality parameters. Another important issue is how to reason with such relationships to draw insightful knowledge about overall data quality. This paper focuses mainly on addressing the last two issues: representational and reasoning issues. To do so, we assume that data quality parameters, such as shown in Table 2, are available for use.

Table 1: Data Quality Indicators

Indicator	data #1	data #2	data #3
Source	DB#1	DB#2	DB#3
Creation-time	6/11/92	6/9/92	6/2/92
Update-frequency	daily	weekly	monthly
Collection-method	bar code	entry clerk	radio freq.

Table 2: Data Quality Parameters

Parameter	data #1	data #2	data #3
Credibility	High	Medium	Medium
Timeliness	High	Low	Low
Accuracy	High	Medium	Medium

The mechanism investigated in this paper is the *data quality reasoner*. This is a simple data quality judgment model based on the notion of a "census of needs." It applies a knowledge-based approach in data quality judgment. The intention is to provide flexibility advantages in dealing with the subjective, decision-analytic nature of data quality judgment. The data quality reasoner provides a framework for representing and reasoning with local relationships among quality parameters to produce an overall data quality level. Such "informating" ability of the data quality reasoner would have significant value for assisting data consumers in judging data quality, particularly when data involved in decision-making come from different, unfamiliar sources.

1.1. Quality Indicators and Quality Parameters

It is worth noting relationship between quality parameters and quality indicators. The essential distinction among quality indicators and quality parameters is that quality indicators are intended (primarily) to represent objective information about the data manufacturing process [Wang & Kon, 1992]. Quality parameters, however, can be user- or application-specific, and are derived from either underlying quality indicators or other quality parameters. The topology of the "quality hierarchy" in this paper is: a single quality parameter being derived from n underlying quality parameters. Each underlying quality parameter, in turn, could be derived from either its underlying quality parameters or quality indicators. For example, a user may conceptualize quality parameter Credibility as one depending on underlying quality parameters such as Source-reputation and Timeliness. The quality parameter Source-reputation, in turn, can be derived from quality indicators such as the number of times that a source supplies obsolete data. This paper assumes that such derivations are complete, and that relevant quality parameter values are available.

1.2. Overview

In general, several quality parameters may be involved in determining overall data quality. This raises the issue of how to specify the degree to which each quality parameter contributes to overall data quality. One approach is to specify the degree, in certain absolute terms, for each quality parameter. It may not, however, be practical to completely specify such values. Rather, people often conceptualize local relationships, such as "Timeliness is more important than the credibility of a source for this data, except when timeless is low." So that, if timeliness is high and Source-credibility is medium, the data may be of high quality. The model presented in this paper provides a formal specification of such local "dominance relationships" between quality parameters.

The issue is, then, how to use these local dominance relationships between quality parameters, and what can be known about data quality from them. Observe that each local relationship between quality parameters specifies the local relative significance of quality parameters. One way to use local dominance relationships would be to rank and enumerate quality parameters in the order of significance implied by local dominance relationships. Finding a total ordering of quality parameters consistent with local relative significance, however, can be computationally intensive. In addition, a

complete enumeration of quality parameters may contain too much information to convey to data consumers any insights about overall data quality. This paper provides a model to help data consumers raise their levels of knowledge about the data they use, and thus make informed decisions. Such a process represents *data quality filtering*.

Our project involves an investigation of a data quality judgment model, with the aim of raising related issues and describing mechanisms behind the use of knowledge about local quality-parameter relationships in data quality judgment. Section 2 discusses a representation for specifying various local relationships between quality parameters. Section 3 discusses the computational component of the quality judgment model. It includes a mechanism for reasoning with local dominance relationships to identify information critical to overall data quality. Finally, Section 4 summarizes this research and suggests future directions for the field of data quality evaluation.

1.3. Related Work

The decision-analytic approach, as summarized in [Keeney & Raiffa, 1976], and utility analysis under multiple objectives, as summarized in [Chankong & Haimes, 1983], describe solution approaches for specifying preferences and resolving multiple objectives. The preference structure of a decision maker or evaluator is specified as a hierarchy of objectives. Through a decomposition of objectives using either subjectively defined mappings or formal utility analyses, the hierarchy can be reduced to an overall value. The decision-analytic approach is generally built around the presupposition of the existence of continuous utility functions. The approach presented in this paper, on the other hand, does not require that dominance relations between quality parameters be continuous functions, or that their interactions be completely specified. It only presupposes that some local dominance relationships between quality parameters exist.

Representational schemes similar to one presented in this paper are investigated, to represent preferences, in sub disciplines of Artificial Intelligence such as Planning [Wellman, 1990,Wellman & Doyle, 1991]. The research effort, however, has focused primarily on issues involved in representing preferences, and much less so on computational mechanisms for reasoning with such knowledge.

2. Data Quality Reasoner

This section discusses the data quality reasoner, called DQR. DQR is a data quality judgment model which derives an overall data quality value for a particular data element, based on the following information:

1) A set, QP, of underlying quality parameters that affect data quality: QP = $\{q_1, q_2, ..., q_n\}$.
2) A set, DR, of local dominance relationships between quality parameters in QP.

In particular, this paper addresses the following fundamental issues that arise in considering the use of local relationships between quality parameters in data quality judgment:

1) How to represent local dominance relationships between quality parameters.
2) What to do with such local dominance relationships.

Section 2.1 presents a representation scheme for specifying local dominance relationships between quality parameters in order to facilitate data quality judgment. Section 2.2 discusses a computational framework which exploits such relationships to draw insights about overall data quality.

2.1. Representation of Local Dominance Relationships

This subsection discusses a representation of local dominance relationships between quality parameters. To facilitate further discussion, additional notations are introduced below. For any quality parameter q_i, let symbol V_i denote the set of values that q_i can take on. In addition, the following notation is used to describe value assignments for quality parameters. For any quality parameter q_i, the value assignment $q_i := v$ (for example, Timeliness := High) represents the instantiation of the value of q_i as v, for some v in V_i. Value assignments for quality parameters, such as $q_i := v$, are called "quality-

parameter value assignments". A quality parameter with a particular value assigned to it is also referred to as an instantiated quality parameter.

For some quality parameters $q_1, q_2, ..., q_n$, for some integer $n \geq 1$, $q_1 \cap q_2 \cap ... \cap q_n$ represents a conjunction of quality parameters. Similarly, $q_1 := v_1 \cap q_2 := v_2 \cap ... \cap q_n := v_n$, for some v_i in V_i, for all i = 1,2, ... , and n, represents a conjunction of quality-parameter value assignments. Note that the symbol $\cap$ used in the above statement denotes the logical conjunction, not set intersection, of events asserted by instantiating quality parameters.

Finally, notation '$\oplus$' is used to state that data quality is affected by quality parameters. It is represented as $\oplus(q_1 \cap q_2 \cap ... \cap q_n)$ that data quality is affected by quality parameters $q_1, q_2, ...,$ and q_n. Statement $\oplus(q_1 \cap q_2 \cap ... \cap q_n)$ is called a *quality-merge statement*, and is read as "the quality merge of q_1, q_2, ..., *and* q_n." Simpler notation, $\oplus(q_1, q_2, ..., q_n)$, is also used. A quality-merge statement is said to be instantiated, if all quality parameters in a quality-merge statement are instantiated to certain values. For example, statement $\oplus(q_1 := v_1 \cap q_2 := v_2 \cap ... \cap q_n := v_n)$ is an instantiated quality-merge statement of $\oplus(q_1, q_2, ..., q_n)$, for some v_i in V_i, for all i = 1, 2, ..., and n.

The following defines a local dominance relationship among quality parameters.

Definition 1 (Dominance relation): Let E_1 and E_2 be two conjunctions of quality-parameter value assignments. E_1 is said to dominate E_2, denoted by $E_1 >_d E_2$, if and only if $\oplus(E_1, E_2, +)$ is reducible to $\oplus(E_1, +)$, where "+" stands for the conjunction of value assignments for the rest of the quality parameters, in QP, which are shown neither in E_1 nor in E_2.

Note that as implied by "+," this definition assumes the context-insensitivity of reduction: $(E_1, E_2, +)$ can be reduced to $\oplus(E_1, +)$, regardless of the values of the quality parameters, in QP, that are not involved in the reduction. Moreover, "+" implies that these uninvolved quality parameters in QP remain unaffected by the application of reduction. For example, consider a quality-merge statement which consists of quality parameters Source-credibility, Interpretability, Timeliness, and more. Suppose that when Source-credibility and Timeliness are High, and Interpretability is Medium, Interpretability dominates the other two. This dominance relationship can be represented as follows:

"Interpretability := Medium $>_d$Source-Credibility := High $\cap$Timeliness := High. "

Then, $\oplus$(Source-credibility := High, Interpretability := Medium, Timeliness := High, +) is reducible to quality-merge statement $\oplus$(Interpretability := Medium, +).

As mentioned at the beginning of Section 2, the evaluation of the overall data quality for a particular data element requires information about a set of quality parameters that play a role in determining the overall quality, QP = $\{q_1, q_2, ..., q_n\}$, and a set DR of local dominance relationships between quality parameters in QP. Information provided in QP is interpreted by DQR as "the overall quality is the result of quality merge of quality parameters $q_1, q_2, ...,$ and q_n, *i.e.*, $\oplus(q_1, q_2, ..., q_n)$." Local dominance relationships in DR are used to derive an overall data quality value. It may be unnecessary or impossible, however, to explicitly state each and every plausible relationship between quality parameters in DR. Assuming incompleteness of preferences in quality parameter relationships, this paper approaches the incompleteness issue with the following default assumption: For any two conjunctions of quality parameters, if no information on dominance relationships between them is available, then they are assumed to be in the indominance relation. The indominance relation is represented as follows:

Definition 2 (Indominance relation): Let E_1 and E_2 be two conjunctions of quality-parameter value assignments. E_1 and E_2 are said to be in the indominance relation, if neither $E_1 >_d E_2$ nor $E_2 >_d E_1$.

When two conjunctions of quality parameters are indominant, a data consumer may specify the result of quality merge of them, according to his or her needs.

2.2. Reasoning Component of DQR

The previous subsection discussed how to represent local relationships between quality parameters. The next question that arises is then how to derive overall data quality from such local dominance relationships, *i.e.*, how to evaluate a quality-merge statement based on such relationships. This task, simply referred to as the "data-quality-estimating problem," is summarized as follows:

> **Data-Quality-Estimating Problem:**
> Let DR be a set of local dominance relationships between quality parameters, $q_1, q_2, \ldots$, and q_n.
> Compute $\oplus(q_1, q_2, \ldots, q_n)$, subject to local dominance relationships in DR.

An instance of the data-quality-estimating problem is represented as a list of a quality-merge statement and a corresponding set of local dominance relationships, *i.e.*, $(\oplus(q_1, q_2, \ldots, q_n), \text{DR})$.

The rest of this section presents a framework for solving the data-quality-estimating problem, based on the notion of "reduction". The following axiom defines the data quality value when only one quality parameter is involved in quality merge.

Axiom 1 (Quality Merge): For any quality-merge statement $\oplus(q_1, q_2, \ldots, q_n)$, if $n = 1$, then the value of $\oplus(q_1, q_2, \ldots, q_n)$ is equal to that of q_1.

Quality-merge statements with more than one quality parameter are reduced to ones with a smaller number of quality parameters. The following define axioms which provide a basis for the reduction. As implied by Definition 1 and the default assumption, any two conjunctions of quality-parameter value-assignments can be in either the dominance relation or in the indominance relation. The following axiom specifies that any two conjunctions cannot be both in the dominance relation and in the indominance relation.

Axiom 2 (Mutual Exclusivity): For any two conjunctions E_1 and E_2 of quality-parameter value assignments, E_1 and E_2 are related to each other in exactly one of the following ways:

1. $E_1 >_d E_2$
2. $E_2 >_d E_1$
3. E_1 and E_2 are in the indominance relation.

The following axiom defines the precedence of the dominance relation over the indominance relation. This implies that while evaluating a quality-merge statement, quality parameters in the dominance relation are considered before those not in the dominance relation.

Axiom 3 (Precedence of $>_d$): The dominance relation takes precedence over the indominance relation.

Reduction-Based Evaluation: A reduction-based evaluation scheme is any evaluation process where the reduction operations take precedence over all other evaluation operations. Definition 1 and axiom 3 allow the reduction-based evaluation strategy to be used to solve the data-quality-estimating problem for quality-merge statements with more than one quality parameter.

The use of dominance relationships to reduce a quality-merge statement raises the issue of which local dominance relationships to apply first, *i.e.*, regarding the order in which local dominance relationships are applied. Unfortunately, the reduction of a quality-merge statement is not always well-defined. In particular, a quality-merge statement can be reduced in more than one way, depending on the order in which the reduction is performed. For example, consider an instance of the data-quality-estimating problem, $(\oplus(q_1, q_2, q_3, q_4, q_5, q_6), \text{DR})$, where DR consists of the following local dominance relationships:

$$q_1 >_d q_2,\ q_2 >_d q_3,\ q_4 >_d q_5,\ q_5 >_d q_6,\ (q_1 \cap q_4) >_d (q_3 \cap q_6),\ (q_2 \cap q_5) >_d (q_1 \cap q_4).$$

Then, the quality-merge statement $\oplus(q_1,q_2,q_3,q_4,q_5,q_6)$ can be reduced to more than one irreducible quality-merge statement. The following show some of them:

- In case that $q_1 >_d q_2$, $q_4 >_d q_5$, and $(q_1 \cap q_4) >_d (q_3 \cap q_6)$ are applied in that order, $\oplus(q_1,q_2,q_3,q_4,q_5,q_6)$ is reducible to $\oplus(q_1,q_4)$, as follows:
  $\oplus(q_1,q_2,q_3,q_4,q_5,q_6)$
  $= \oplus(q_1,q_3,q_4,q_5,q_6)$, by applying $q_1 >_d q_2$,
  $= \oplus(q_1,q_3,q_4,q_6)$, by applying $q_4 >_d q_5$,
  $= \oplus(q_1,q_4)$, by applying $(q_1 \cap q_4) >_d (q_3 \cap q_6)$.
- In case that $q_2 >_d q_3$, $q_5 >_d q_6$, and $(q_2 \cap q_5) >_d (q_1 \cap q_4)$ are applied in that order, $\oplus(q_1,q_2,q_3,q_4,q_5,q_6)$ is reducible to $\oplus(q_2,q_5)$, as follows:
  $\oplus(q_1,q_2,q_3,q_4,q_5,q_6)$
  $= \oplus(q_1,q_2,q_4,q_5,q_6)$, by applying $q_2 >_d q_3$,
  $= \oplus(q_1,q_2,q_4,q_5)$, by applying $q_5 >_d q_6$,
  $= \oplus(q_2,q_5)$, by applying $(q_2 \cap q_5) >_d (q_1 \cap q_4)$.

As illustrated in this simple example, the reduction of a quality-merge statement is not always well-defined.

3. First-Order Data Quality Reasoner

This section investigates a simpler data quality reasoner that guarantees the well-defined reduction of quality-merge statements, by making certain simplifying assumptions. To facilitate the next step of derivation, an additional definition is introduced.

Definition 3 (First-Order Dominance Relation): For any two conjunctions E_1 and E_2 of quality parameters such that E_1 and E_2 are in the dominance relation, E_1 and E_2 are said to be in the first-order dominance relation, if each of E_1 and E_2 consists of one and only one quality parameter.

The first-order data quality reasoner, in short called DQR1, is a data quality judgment model that satisfies the following:

First-order Data Quality Reasoner (DQR1)
1. Axioms 1, 2, and 3 hold.
2. Only indominance and first-order dominance relationships are allowed.
3. $<_d$ is transitive (*i.e.*, transitive dominance relation).

In the first-order data quality reasoner, higher-order dominance relationships, such as $q_1 \cap q_2 \cap q_3 >_d q_4$ or $q_4 \cap q_5 >_d q_1 \cap q_2$, are not allowed. In addition, the first-order data quality reasoner requires that the dominance relation be transitive. This implies that for any conjunctions of quality-parameter value assignments, E_1, E_2, and E_3, if $E_1 >_d E_2$ and $E_2 >_d E_3$, then $E_1 >_d E_3$. Transitivity of the dominance relation implies the need for an algorithm to verify that, when presented with an instance of the quality-estimating problem $(\oplus(q_1,q_2,\ldots,q_n), DR)$, dominance relationships in DR do not conflict with each other. Well-known graph algorithms can be used for performing this check (T H Cormen, Leiserson, & Rivest, 1990).

Quality-merge statements can be classified into groups, with respect to levels of the reducibility, as defined below.

Definition 4 (Irreducible Quality-Merge Statement): For any instantiated quality-merge statement $e=\oplus(q_1:=v_1,q_2:=v_2,\ldots,q_n:=v_n)$ such that $n \geq 2$, for some v_i in V_i, $\forall i = 1,2,\ldots,$ and n, e is said to be irreducible, if for any pair of quality-parameter value assignments in e, say $q_i:=v_i$ and $q_j:=v_j$, $q_i:=v_i$ and $q_j:=v_j$ are in the indominance relation. Similarly, any quality-merge statement which consists of one and only one quality parameter is said to be irreducible.

Definition 5 (Completely-Reducible Quality-Merge Statement): For any instantiated quality-merge statement $e=\oplus(q_1:=v_1,q_2:=v_2,...,q_n:=v_n)$ such that $n\geq 2$, for some v_i in V_i, $^\forall i$ = 1,2,..., and n, e is said to be completely reducible, if for any pair of quality-parameter value assignments in e, say $q_i:=v_i$ and $q_j:=v_j$, $q_i:=v_i$ and $q_j:=v_j$ are in the dominance relation.

The next two sub-sections discuss algorithms for evaluating a quality merge statement in DQR1. This process is diagrammed in Figure 1. Algorithm Q-Merge is the top level algorithm which receives as input a quality-merge statement and the corresponding quality parameter relationship set. Within Algorithm Q-Merge, there is a two stage process. First, the given quality-merge statement is instantiated accordingly. It then calls Algorithm Q-Reduction to reduce the QMS into its corresponding irreducible form.

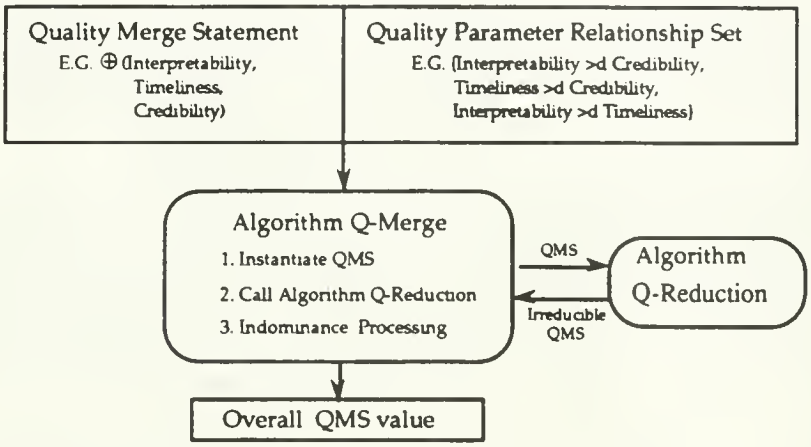


Figure 1. The Quality-Merge Statement (QMS) Evaluation Process

3.1. Reduction of Quality-Merge Statement: Algorithm Q-Reduction

This section describes an algorithm, called Q-Reduction, for reducing a quality-merge statement into an irreducible quality-merge statement, according to local dominance relationships between quality parameters. This subsection continues to assume that all dominance relationships are first-order.

Algorithm Q-Reduction
Input: (e, DR),
where $e = \oplus(q_1:=v_1,q_2:=v_2,...,q_n:=v_n)$ for some v_i in V_i, $^\forall i$ = 1,2,..., and n, and
DR is a set of local dominance relationships between quality parameters $q_1,q_2,...,$ and q_n.
Output: An irreducible quality-merge statement for e.

Let Ω be the set of the quality-parameter value assignments in e: $\Omega = \{q_1:=v_1,q_2:=v_2,...,q_n:=v_n\}$.
1. LOOP for each local dominance expression, say $q_i:=a_i >_d q_j:=a_j$, in DR,
2. IF ($q_i:=a_i \in \Omega$) and ($q_j:=a_j \in \Omega$) THEN $\Omega \leftarrow \Omega - \{q_j:=a_j\}$
;; The irreducible QMS consists of the quality parameters in the modified Ω which the loop results in.
;; Let Ω' denote the final modified Ω. Then, e is reducible to e' which consists of the quality parameters in Ω'.
3. Return $\oplus(\Omega')$

Figure 2: Algorithm Q-Reduction

Algorithm Q-Reduction in Figure 2 takes as input an instantiated quality-merge statement e and DR, and returns as output an irreducible quality-merge statement of e. The instantiated quality-merge statement e is reduced as follows.

For expository purposes, suppose that $e = \oplus(q_1:=v_1, q_2:=v_2, \ldots, q_n:=v_n)$, for some v_i in V_i, for all i = 1,2,... , and n, and let Ω be a dynamic set of quality-parameter value assignments, which is initialized to $\{q_1:=v_1, q_2:=v_2, \ldots, q_n:=v_n\}$. For any pair of quality-parameter value assignments $q_i:=v_i$ and $q_j:=v_j$ in e, if $q_i:=v_i >_d q_j:=v_j$ is a member of DR, then e is reducible to a quality-merge statement with the quality parameters in Ω less $q_j:=v_j$, by Definition 1. This allows removing $q_j:=v_j$ from Ω, if both $q_i:=v_i$ and $q_j:=v_j$ are elements in Ω. Continue the process of removing dominated quality parameters, until no pair of the quality parameters in Ω are related in the dominance relation. Let Ω' denote the modified Ω produced at the end of this removal process. The quality merge of the quality parameters in Ω' is the corresponding irreducible quality-merge statement of e, and the algorithm returns $\oplus(\Omega')$. It is proven in (Jang & Wang, 1991) that Algorithm Q-Reduction shown in Figure 2 always results in a unique output in the first-order data-quality reasoner, in that all dominance relations must be first-order.

3.2. Algorithm Q-Merge

When presented with an instance of the quality-estimating problem $(\oplus(q_1, q_2, \ldots, q_n)$, DR) for some integer n, Algorithm Q-Merge first instantiates the given quality-merge statement, accordingly. The instantiated quality-merge statement is then reduced until the reduction process results in another instantiated quality-merge statement which cannot be reduced any further (using Q-Reduction). This raises the issue of how to evaluate an irreducible quality-merge statement.

Unfortunately, the evaluation of an irreducible quality-merge statement is not always well-defined. When evaluating an irreducible quality-merge statement, the number of orders in which the quality merge operation can be applied grows exponentially with the number of quality parameters in the statement. In particular, certain quality-merge statements may be merged in more than one way, depending on the order in which the merge is performed. It is possible that this set might include every element of V_i's. This paper evades this problem by presenting quality-parameter value assignments in the irreducible quality-merge statement returned by Algorithm Q-Reduction so that a user may use this information presented, according to his or her needs. Figure 3 summarizes Algorithm Q-Merge.

Algorithm Q-Merge
Input: (e, DR),
where $e = \oplus(q_1, q_2, \ldots, q_n)$, for some integer n and
DR is a set of local dominance relationships between quality parameters $q_1, q_2, \ldots$, and q_n.
Output: Overall data quality value produced by evaluating e.

1. Instantiate e.
   ;; Suppose that $q_1, q_2, \ldots$, and q_n are instantiated as $v_1, v_2, \ldots$, and v_n, respectively, for some v_i in V_i
   ;; for all i = 1,2, ..., and n.
2. $e' \leftarrow$ IF (n = 1) THEN e
   ELSE Q-Reduction($\oplus(q_1:=v_1, q_2:=v_2, \ldots, q_n:=v_n)$, DR)
3. Present quality-value assignments in e'.

Figure 3: Algorithm Q-Merge

4. Discussion

We have presented a knowledge-based framework for data quality judgment that: (1) allows specifying local dominance relationships between quality parameters, (2) performs the reduction of quality-merge statements, and (3) derives a value for overall data quality. A knowledge-based approach was applied to data quality judgment, to provide significant flexibility advantages in representing data-consumer-specific requirements on data, and thereby tailoring data quality judgment to data consumers' needs.

In addition, our analysis has identified issues that must be addressed in order for the quality judgment model presented in this paper to be of practical use. The rest of this section considers the

limitations of the approach explored in this paper, and suggests future directions for the field of data quality judgment.

Higher-order data quality reasoner: The problem associated with the reduction of quality-merge statements was discussed in Section 2.2. The first-order data quality reasoner evades the problem of ill-defined reduction by prohibiting higher order relationships. Real-world problems, however, often involve more complex relationships than first-order relationships between quality parameters. In order to deal with higher-order relationships, both the representational and algorithmic components of the first-order data quality reasoner would need to be extended.

Data Acquisition/A hierarchy of quality indicators and parameters: This research assumed that values of quality parameters are available so that quality-merge statements can be instantiated properly. Issues of how to represent and how to streamline such values to the data quality reasoner, however, must be addressed. One approach to these issues would be to organize quality parameters and quality indicators in a hierarchy. Then, to each data element or type can be attached information about how to compute a value of a quality parameter. Such a hierarchy would allow the derivation of a quality parameter value from its underlying quality parameters and quality indicators. A tool for automatically constructing such a hierarchy and computing quality-parameter values would enhance the utility of the data quality reasoner.

Knowledge Acquisition: The capability of using local dominance relationships, which are typically user- or application-specific, allows us to build systems more adaptable to customers' needs. As application domains are complex, however, it becomes increasingly difficult to state all the relationships that must be known. Such knowledge acquisition bottlenecks could be alleviated through development of a computer program for guiding the process of acquiring relationships between quality parameters.

User interface for cooperative problem solving: As mentioned in Section 3, the evaluation of an irreducible quality-merge statement is not well-defined. Different orders in which quality parameters in an irreducible quality-merge statement are evaluated may result in different values. This research dealt with the need to evaluate an irreducible quality-merge statement by simply presenting information on quality parameters in an irreducible quality-merge statement. Development of a user interface which allows evaluating irreducible quality-merge cooperatively with a data consumer could lessen the problem.

In the continuous cycle of measurement, analysis, and improvement for data quality management, it is crucial that a methodology be developed for judging data quality. In particular, while each individual data supplier may maintain integrity and consistency of its own data, such local integrity and consistency do not necessarily guarantee that data from different suppliers display the same level of quality. The development of a system that can assist data consumers in judging if data meets their requirements is important, particularly when decision-making involves data from different, foreign sources. The model presented in this paper provides a first step toward such a system.

Acknowledgments

The authors would like to thank Peter Szolovits and Stuart Madnick for their support, members of Composite Information Systems Laboratory for their comments, and Alexander Ishii for his numerous insightful suggestions on earlier versions of this paper.

References

[1] Bonoma, T. V. (1985). Case research in marketing: opportunities, problems, and a process. *Journal of Marketing Research*, **22**, pp. 199-208.

[2] Burnham, D. (1985). *FBI Says 12,000 Faulty Reports on Suspects are Issued Each Day*. NY.

[3] Chankong, V. & Haimes, Y. Y. (1983). *Multiobjective Decision Making: Theory and Methodology* . New York, N.Y.: Elsevier Science.

[4] Jang, Y. & Wang, Y. R. (1991). *Data Quality Calculus: A data-consumer-based approach to delivering quality data*. (CISL-91-08) Composite Information Systems Laboratory, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139 November 1991.

[5] Johnson, J. R. (1990). Hallmark's Formula For Quality. *Datamation*, , pp. 119-122.

[6] Keeney, R. L. & Raiffa, H. (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs* . New York: John Wiley & Son.

[7] Laudon, K. C. (1986). Data Quality and Due Process in Large Interorganizational Record Systems. *Communications of the ACM*, **29**(1), pp. 4-11.

[8] T H Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1990). *Introduction to Algorithms* . Cambridge, MA, USA: MIT Press.

[9] Wang, R. Y. & Kon, H. B. (1992). *Toward Quality Data: An Attributes-based Approach to Data Quality*. (CISL-92-04) June 1992.

[10] Wang, Y. R. & Guarrascio, L. M. (1991). *Dimensions of Data Quality: Beyond Accuracy*. (CISL-91-06) Composite Information Systems Laboratory, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139 June 1991.

[11] Wellman, M. P. (1990). *Formulation of Tradeoffs in Planning Under Uncertainty* . Pitman and Morgan Kaufmann.

[12] Wellman, M. P. & Doyle, J. (1991). *Preferential Semantics for Goals*. The Proceedings of the 9th National Conference on Artificial Intelligence, 1991. pp.