ENHANCING DATA QUALITY THROUGH ATTRIBUTE-BASED METADATA AND COST EVALUATION IN DATA WAREHOUSE ENVIRONMENTS

Yu-Chi Chu[1] Shan-Shan Yang[2] Chen-Chau Yang[1]*
[1]*Department of Electronic Engineering*
National Taiwan University of Science and Technology
Taipei, Taiwan 106, R.O.C.
[2]*Program Management & Coordination*
Telecommunication Laboratories
Chunghwa Telecom Co., Ltd.
Taipei, Taiwan 106, R.O.C.

Key Words: data warehouse, data quality, metadata, cost/benefit evaluation model.

ABSTRACT

Data quality will be a significant issue as data warehousing becomes more and more popular. This paper aims at investigating and analyzing the data quality issues in data warehouse environments. We present an attribute-based metadata model for identifying data quality. A four-phase process is introduced for data quality management during the life cycle of data warehouses. Overall data quality conditions can be identified and related information can be provided for determining whether the data meet "fit to use" criteria and whether they need to be improved. Furthermore, we use a cost/benefit evaluation model to ferret out the poor-quality data and set priorities for improvement given limited resources. Our approach allows system developers to document relevant quality data as metadata, which may be associated with the whole life cycle of data warehouses. Quality metadata not only can enrich the interpretation of attribute data, but can also provide diagnostic information for finding the reasons for and the sources of errors. In addition, the cost/benefit evaluation model developed may provide a foundation for the quantitative analysis of data quality.

I. INTRODUCTION

In recent years, data warehouse systems (DWS) have attracted a great deal of interest in both academic and industrial communities. In the typical data warehouse architecture, the data subject to analysis is integrated from multiple sources, both internal and external, and selected information is extracted in advance and stored in a repository. Generally, the information stored in the warehouse can be structured and organized in a form that makes it easy to use for applications. A data warehouse can therefore be seen

*Correspondence addressee

as a set of materialized views defined over the remote sources, and warehoused data is usually used for decision making, rather than for operations.

There are some problems that should be addressed in data warehousing (Garcia-Molina, *et al.*, 1999, Kimball, 1996). For example, data from different sources may have serious semantic differences, and is likely to contain syntactic inconsistencies. Moreover, the desired data may simply not have been gathered. Therefore, data warehousing projects may not succeed for various reasons, such as poor system architecture or unacceptable query performance, but nothing is more certain to yield failure than lack of concern for the issue of data quality. If the data in the warehouse do not meet quality characteristics required to support decisions, the data warehouse effort will be blamed for the shortcomings. Poor-quality data will lead either to wrong decisions being made, or knowledge workers losing confidence in the data warehouse. Although some companies recently have become aware of the importance of high-quality data and some straightforward approaches have been proposed for data quality management, there is a definite need for comprehensive approaches to improve data quality.

On the basis of the consideration of quality assurance (QA), we propose a methodology for data quality assurance in data warehouse environments. Our methodology not only concerns the processes of improving data quality, but also takes into account the cost of data quality improvement. Since metadata plays an important role throughout the life cycle of data warehouses, we adopt an attribute-based metadata model for identifying data quality. By using the information provided by the quality metadata model, we can identify the sources of poor-quality data. Furthermore, a cost/benefit model is used to identify the most fatal poor data and set the priorities for improvement, given limited resources.

This paper is organized as follows. Section II presents a framework for data quality assurance and how data quality can be represented as a hierarchical structure, as well as a four-phase process for data quality management (DQM). The proposed attribute-based metadata model is discussed in Section III. Section IV describes the evaluation of data quality based on the cost/benefit model. Section V reviews related work, while Section VI contains concluding remarks and future research directions.

II. FRAMEWORK OF DATA QUALITY ASSURANCE

Data quality has two distinct aspects: one is the "correctness" of data (such as accuracy and consistency), and the other involves the appropriateness

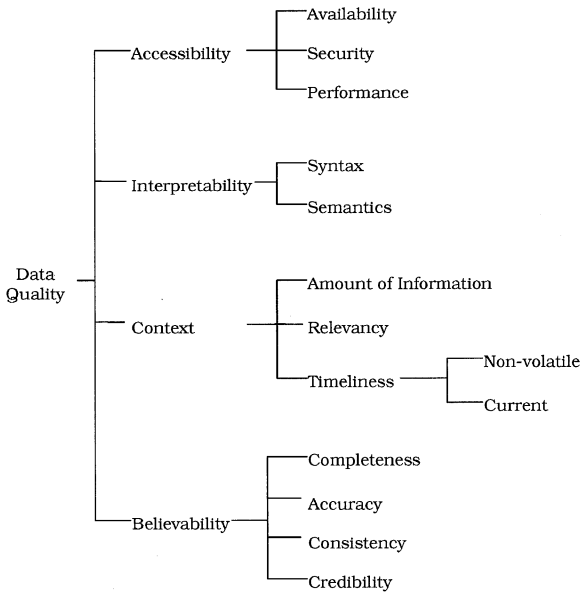

Fig. 1 Data quality hierarchy

of data for some intended purposes. Data producers and users generally assume that the purpose of data quality assurance is to provide the best data possible. However, this obscures the need to evaluate data. The implication is that if a data set is the best available and is as good as it can be made, then there are no other options than to use it. In this case, there is no point in worrying about just how good it can be made. The flaw in this is that merely saying that a data set is as good as it can be made does not tell us *how* good it is or whether it is *any* good at all. What may be considered good data in one case may not be sufficient in another case. For example, an analysis of the financial position of a firm may require data in units of thousands of dollars while an audit requires precision to the cent. Therefore, the term "data quality" may best be defined as "fit to use," which implies the quality of the data in the warehouses is accurate enough, timely enough, and consistent enough for the organization to make reasonable decisions (Orr, 1998; Wand and Wang, 1996).

1. Data Quality Hierarchy

On the basis of the goal of "fit to use," data quality can be classified into four dimensions, and each dimension may be composed of several "data quality factors." Moreover, each data quality factor may have some "data quality indicators." Therefore, we can form a hierarchical structure of data quality for investigating the relationship between each pair of levels in a systematic approach. Fig. 1 shows the hierarchical structure of data quality. The meanings of

the four dimensions of data quality are briefly discussed as follows.

(i) Accessibility

From the user's point of view, a DWS should provide an efficient mechanism for accessing the data in the data warehouse under certain considerations of security. The mechanism should be able to reduce the effort of searching in a large and poorly structured information space, as well as avoiding interference of data analysis with operational data processing. When the amount of data in the warehouse becomes huge, the factor of performance should be taken into account for evaluating the balance between access efficiency and system loading.

(ii) Interpretability

It remains difficult for DWS users to interpret the data if the semantics of data description languages for data warehouse schemata is weak, fails to take domain-specific aspects into account, and not formally defined. The data interpretability dimension is concerned with data description, such as data layout for legacy systems and external data, table description for relational databases, primary and foreign keys, aliases, defaults, domains, explanation of coded values, etc.

(iii) Context

We adopt the amount of information, relevancy, and timeliness as three factors for evaluating the data quality of context. A great deal of information might help the process of decision making, but obviously it might also cause the degradation of system performance and waste of resources. Thus, the relevance between user's needs and the data in the warehouse should be evaluated during the construction of the data warehouse. With regard to the factor of timeliness, we can evaluate it by examining two indicators: *non-volatile*, which means the use of data is independent of temporal relationships, and current, which means dependency on temporal relationships.

(iv) Believability

Since most DWS users often do not know the credibility of the sources and the accuracy of the data, the believability of data is hampered. In addition, schema languages are too weak to ensure completeness and consistency testing. We can evaluate the believability of data in terms of the completeness, consistency, accuracy, and credibility (Ballou and Pazer, 1987; Huh *et al.* 1990). Completeness means the percentage of the real-world information entered in the sources and/or the warehouse. For example, completeness could rate the extent to which a string describing an address did actually fit in the size of the attribute, which represents the address. The accuracy stands for the correctness of the data entry process, which happened at the source. The consistency represents the logical harmony of the information, both in syntactic and semantic aspects. The credibility describes the trustworthiness of the sources that provided the information.

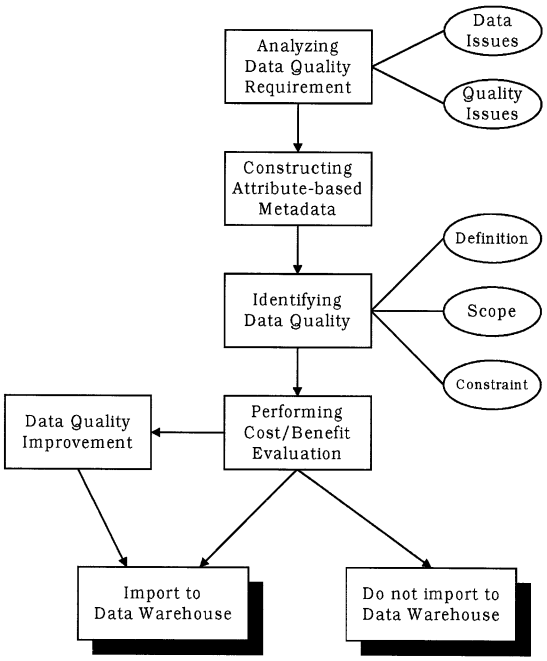

Fig. 2 The process of data quality management

2. Process of Data Quality Management

The data quality in the warehouse is determined not just by a single process; all the processes that take place in the warehouse environment may affect it. Thus, quality considerations have accompanied data warehouse research from the beginning. Generally, data stored in the warehouse come from various sources including internal databases and external data resources. If there are quality problems in those sources, these problems will be moved to the warehouse accordingly. This will cause an unpredictable situation when a warehouse is used for decision making. Furthermore, even if there are no quality problems in those data sources, data errors or the degradation of data quality might occur during the processes of data integration and transformation when constructing and maintaining data warehouses. Hence, there is a need to develop a mechanism or operational procedure that can be used to ensure the data quality during the life cycle of data warehouses.

In this paper, we propose a systematic approach for data quality assurance. In this approach, data quality management consists of the following four phases as shown in Fig. 2. After accomplishing these activities, we may determine the data items that need to be improved to meet the goal of "fit to use."

Phase 1: Analyzing data quality requirements. This phase is similar to the logical design of conventional database systems, wherein the system designers have to figure out semantic ambiguities and syntactic inconsistencies from various sources. Data issues and quality issues should be taken into account during this phase. For determining the type and number of data quality factors to be covered to meet user's needs, the results of this phase will be the specifications for data quality management requirements in the data warehouse.

Phase 2: Constructing attribute-based metadata. Data warehouse systems usually have a multi-dimensional schema to store integrated data from different sources. To ensure the data quality, we add an extra dimension dedicated to the description of data quality for each specific attribute. Moreover, the quality data can be combined with attribute data to simplify the description. Although the description of data quality will cause overhead in storage resources, we believe the benefits of having good data quality can cover the cost of storage space.

Phase 3: Identifying data quality. A data warehouse may support decision making for users at different levels in an organization. The corresponding requirements of data quality are different for each user. This is consistent with the principle of "fit to use." For example, for the quality factor of timeliness, some users may need data collected during the last year, while the others may need the data collected during the past decade for detailed analysis. Therefore, we need to identify data that may not fit needs and the causes of the mismatches.

Phase 4: Performing cost/benefit evaluation. Once the unqualified data are identified, we have to find out how to improve the quality of those data. In practice, we need to take cost issues into account to determine the efforts needed to improve the quality of unqualified data. Since it is impractical to achieve a flawless state of data quality, we need to perform a cost and benefit analysis to determine to which level we should improve the quality of unqualified data.

After finishing the above processes, the results may provide the system designer with helpful support to adopt appropriate strategies for data quality assurance. We may import data that meet the requirements of data quality, to the warehouse immediately. Unqualified data can be divided into two categories. The first portion will be imported to the warehouse after we improve the quality, for the cost of improvement is acceptable. Another portion represents unqualified data which are too expensive to be improved. It is a tradeoff issue whether such data should be imported to the warehouse, and the system designers and users should make decisions with deliberation.

III. ATTRIBUTE-BASED METADATA MODEL

1. Metadata to Support Data Quality

Within the data warehousing architecture, users do not entirely control the quality of the data they use. When a data set is obtained from a data source or an intermediate data center, its quality (accuracy, consistency, etc.) may already have been determined. This implies that multiple evaluations must be supported, and the results of each evaluation must be recorded for use by future data users and by system developers and maintainers. Finally, it is vital to consider the ways that data may be transformed between initial production and final use. Various agents and users themselves may combine, aggregate, filter, edit and modify data from different sources in order to prepare a data set for a specific use. These transformation processes–along with the processes that generate data initially–all affect the quality of the resultant data. Therefore, in addition to recording information about data values themselves and evaluating the quality of these values, it is important to record information about the processes that affect data, both to ensure that data quality is not corrupted and to allow improving these processes.

There are many kinds of metadata that can be associated with a data warehouse, including metadata to restrict access to data, to facilitate sharing and interoperability, to characterize and index data, etc. Metadata may also be used to define the user's expectations of data quality and to describe the conditions of data quality in the data warehouses. Thus, the metadata may serve as a data quality profile, which can be easily extended as required (Rothenberg, 1996).

We augment an attribute-based data quality model in (Wang *et al.*, 1995a) with more flexible and simplified considerations. Since most data warehouse systems have multi-dimensional characteristics for storing and retrieving data, we may dedicate one of the dimensions to specifying quality metadata. Therefore, data quality can be linked to a specific data attribute by an internal linkage. For example, Table 1 shows an extended table in which the original data table is made up of < ID, Addr, Owner, Profit_est. >. After analyzing the requirements for data quality, we add three items of quality metadata, < Entry_date,

Table 1 Attributed-based metadata model with quality description

ID	Addr.	Owner	Profit_est.	Entry_date	Evaluator	Entry
B001	Taipei	David	$1,000	1998-03-21	Monica	Mary
B002	Taichung	Mike	$3,000	1998-03-21	Bill	John
⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Data attribute			Quality metadata		

Table 2 Extended attribute with quality description

ID	Addr.	Owner	Profit_est. Δ
B001	Taipei	David	$1,000
B002	Taichung	Mike	$3,000
⋮	⋮	⋮	⋮

Table 3 Level one of quality metadata

Profit_est. Δ	Source Δ	Entry
B001Δ	Monica	Mary
B002Δ	Bill	John

Table 4 Level two of data quality metadata

Source Δ	Evaluator	Entry_date
B001 Δ	Monica	1999-03-21
B002 Δ	Bill	1999-03-21

Evaluator, Entry > for the attribute of Profit_est. These metadata represent the date of data being input, who evaluated the profit, as well as who gathered the data.

2. Understanding Quality with Metadata

For ensuring proper association of data attributes and data quality in the warehouse, the results obtained from the phase of requirement analysis will provide helpful support during implementation. Since each data attribute may have several quality metadata items, we need to develop a mechanism for investigating their association relationship within the schemas of data warehouse systems. If a data attribute connects with quality metadata, then a symbol of "Δ" will be used to identify all the data belonging to the attribute which can be associated with its quality metadata.

The overall picture of the implementation forms a multi-level framework. Quality metadata at the same level are viewed as a *quality schema*. The primary index in the schema is a *quality index* which connects the attribute data and quality metadata. For example, if we consider the data quality of profit estimation, such as, who made the estimation and when the data were collected, we may add the symbol "Δ" to the attribute of <Profit_est.>. This will form a quality schema as shown in Table 2.

The attribute <Profit_est. Δ> in Table 2 can be further extended with its quality metadata to form a new quality schema including quality description as shown in Table 3; the attribute <Source> in Table 3 can make further descriptions regarding the quality of data sources shown as Table 4. In the example, we may find that the profit estimation for certain stores depends on the *evaluator's quality*. The decision-maker may judge the believability of the data in the warehouse based on the person who evaluated the data and when the data were collected. A detailed illustration of the implementation of quality metadata can be found in (Yang, 1999).

Our approach for constructing the attribute-based metadata differs from the approach in (Wang 1995a) in two ways. First, we adopt a multi-level framework to implement a flexible mechanism for data quality assurance. Secondly, our approach may benefit the process of data quality query, and users are able to query the data quality without modification of the query language. Therefore, the system for assuring data quality in the warehouse can be improved overall. Furthermore, when data quality issues are associated with the life cycle of data warehouses, we need to take integration issues into account. As regards an inconsistent situation during the process of data access, we have to modify quality metadata simultaneously when the corresponding attribute data are modified or deleted.

IV. IDENTIFYING AND EVALUATING DATA QUALITY

In this section, we present the procedures of data quality identification and cost/benefit evaluation. First, we will discuss how to establish data quality constraints in conjunction with attribute-based metadata, and how to detect errors and anomalies. Secondly, we will present a cost/benefit evaluation model to identify priorities for improving poor-quality data.

1. Data Quality Identification

In data warehouse environments, data quality identification concerns not only the correctness of data in the warehouse, but also the characteristics of data format, syntax, and semantics as well as data consistency with the data sources. Thus, the procedures of data quality identification of data warehouses are more complicated than traditional databases. However, according to the surveys in (Barquin and Edelstein, 1997; Parsaye and Chignell, 1993), almost 80% of data quality problems are caused by 20% of defects. We may refer to such a situation as the Pareto principle, which implies that we should focus on the problems and causes that have the biggest impact on quality and cost. On the basis of quality requirements that are defined in terms of data quality metadata, we can establish a list of data quality constraints based on the quality dimension of context introduced in Section II. For example, some data items might have expiration date considerations. If the profit estimation in Table 1 was made two years ago, it might not reflect the current situation at all.

By using data quality constraints, data errors and anomalies may be detected based on rules that are constructed in advance. For example, in the following, constraint #1 states that attribute data Entry_date has a constraint on expiration date; constraint #2 states that when the attribute data Profit_est. is greater than $3000, and metadata shows that the data is from the source named "Monica," such data may be regarded as an anomaly because it contradicts user's expectations.

%Constraint #1:
IF (M.Entry_date<1999-03-31)
THEN Condition = expired
%Constraint #2:
IF (A.Profit_est.>3000) AND (M.Source ="Monica")
THEN Condition = anomaly

Considering the balance of cost and benefit, we need to rank the data quality indicators based on current data. The purpose of ranking is to assist the warehouse's managers to find out the most critical source that causes poor quality and affects the data quality in the data warehouse. Different subjects and concerns may be used to classify ranking criteria. For example, the rate of data errors, the cost caused by defective data and how many resources are required to fix poor-quality data.

The outcomes of data quality identification provide a foundation for further cost/benefit evaluation. We are currently implementing a set of tools for assisting the warehouse's managers to generate and maintain data quality constraints, as well as for detecting errors and anomalies. These tools will consist of a group of graphic user interfaces and middleware used to access the attribute data and quality metadata in warehouses. A more detailed description of the implementation of data quality identification can be found in (Yang 1999).

2. Cost/benefit Evaluation

The real difficulty with data quality is change (Orr, 1998). The data in the data warehouse is static, but the real world keeps changing. Even if data is 100% in agreement with the real world at time t_0, at t_1 it will be slightly off, and at t_2 it will be even further off. As data warehouses get older their data quality problems tend to worsen. Therefore, the relationship between time and data quality may provide a foundation for the quantitative analysis of data quality.

We define the degradation of data quality as a function of time, denoted as $Q(t)$, representing the data quality in a data warehouse at a certain time point t. In addition, the data quality are composed of several indicators; thus $Q(t)$ can be defined as

$$Q(t) = \sum_{i=0}^{n} Q_i(t) \tag{1}$$

where $Q_i(t)$ represents the degradation of each data quality indicator. According to Eq. (1), we may not be able to find out the priority for each indicator, and all indicators carry the same weight of cost and importance. Yet if we view the degradation as the proportions of poor-quality data in warehouses, $Q(t)$ can be quantified to a real number between 0 ~ 1, then Eq. (1) can be modified as follows.

$$Q(t) = \sum_{i=1}^{n} \frac{\rho_i}{W} \tag{2}$$

where W represents the amount of attribute data, and ρ stands for the number of poor-quality data.

We propose that the total cost of data quality should include the *lost cost* and the *improvement cost*. In accordance with these two issues, we can determine what kinds of data items in the data warehouse should be modified and improved. We define the lost cost and the improvement cost as follows.

Lost cost means the cost caused by poor-quality data. The cost may be expressed in terms of lost funding, lost production, lost assets or legal liability. Generally, lost cost is dependent on the degradation of data quality.

Improvement cost means the cost to improve data quality to a certain level. It is dependent on the number of data quality indicators that need to be improved or modified.

Let poor-quality data exist from $t=t_0$, the

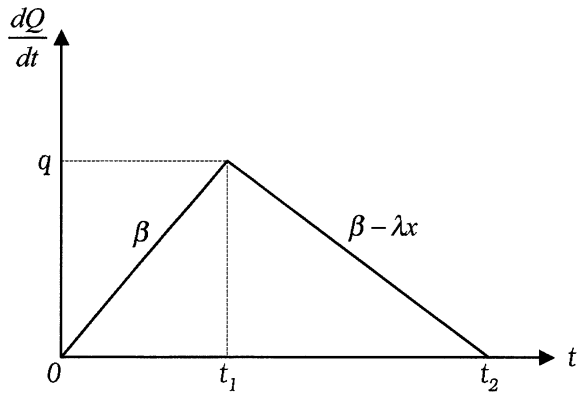

Fig. 3 $\frac{dQ}{dt}$~t relationship

improvement activities start at $t=t_1$ and finish at $t=t_2$. Then, the lost cost caused by poor-quality data is $Q(t_2)$, and the improvement cost is $Q(t_2-t_1)+C$, where C is the cost of time-independent issues such as material resources. Moreover, we found that $\frac{dQ}{dt}$ is more convenient for evaluation than $Q(t)$, because $\frac{dQ}{dt}$ represents the proportions of poor-quality data within a time unit, and appropriately stands for the degradation of data quality in the warehouse. We therefore construct an evaluation model based upon the following criteria, i.e. modeling $\frac{dQ}{dt}$~t for its distributed relationship.

(i) Let the degradation of data quality $\frac{dQ}{dt}$ be the scale of poor-quality data in the data warehouse within a time unit.

(ii) The lost cost is directly proportional to $\frac{dQ}{dt}$ with two coefficients C_1, C_2.

(iii) $\frac{dQ}{dt}$ is directly proportional to time t, and there is a coefficient β to identify the rate of the degradation of data quality .

(iv) When $t \geq t_1$, the rate of the degradation of data quality becomes $\beta-\lambda x$, where λ is the average rate of the improvement. In an ideal condition, the assumption should satisfy $\beta<\lambda x$.

(v) On the basis of the temporal relationship, the improvement cost of data quality can be classified as follows.

- **time-dependent**: let C_2 be the coefficient of improvement cost within a time unit, then the improvement cost of each quality indicator is $C_2(t_2-t_1)$, for instance, the input of man-hours can be viewed as a time-dependent cost.
- **time-independent**: let C_3 be the coefficient of improvement cost, for instance, the input of material resources is a time-independent cost.

(vi) We classify poor-quality data based on each data quality indicator, therefore $\frac{dQ}{dt}$ can be determined by $\sum \frac{dQ_i}{dt}$ (i=1, 2, ..., n), and we obtain

$$\frac{dQ(t)}{dt} = \frac{d\sum_{i=1}^{n} Q_i(t)}{dt} = \sum_{i=1}^{n} \frac{dQ_i}{dt} \tag{3}$$

We construct the model based upon the $\frac{dQ}{dt}$~t elationship as shown in Fig. 3 using the assumptions mentioned above. During $t_0 \leq t \leq t_1$, the degradation rate of data quality, denoted as β, will be a linear ratio along with time. As $t=t_1$, the amount of inadequate data become $\frac{dQ(t=t_1)}{dt}=q$.

When $t_0 \leq t \leq t_2$, is the time period when inadequate data occurs, and the degradation of data quality is $Q(t_2)=\int_0^{t_2} \frac{dQ}{dt}dt$, i.e. the triangular area in Fig. 3 when $0 \leq t \leq t_2$. We compute the lost cost **LC** based on assumption 2 introduced earlier:

$$\boldsymbol{LC} = C_1 Q(t_2) = C_1 \int_0^{t_2} \frac{dQ}{dt}dt = \frac{1}{2} C_1 q t_2 \tag{4}$$

Let $\frac{q}{t_2 - t_1}=\lambda x-\beta$, we obtain

$$\boldsymbol{LC} = \frac{1}{2} C_1 q t_1 + \frac{C_1 q^2}{2(\lambda x - \beta)} \tag{5}$$

During $t_1 \leq t \leq t_2$, the time period that covers improvement activities; the degradation of data quality becomes $Q(t_2-t_1)=\int_{t_1}^{t_2} \frac{dQ}{dt}dt$, i.e. the triangular area in Fig. 3 when $t_1 \leq t \leq t_2$. Based on assumption 5 introduced earlier, we obtain the improvement cost as follows.

$$\boldsymbol{IC} = C_2 x(t_2 - t_1) + C_3 x = \frac{C_2 q x}{\lambda x - \beta} + C_3 x \tag{6}$$

Accordingly, adding lost cost and improvement cost, we have

$$\boldsymbol{TC} = \frac{1}{2} C_1 q t_1 + C_3 x + \frac{C_1 q^2}{2(\lambda x - \beta)} + \frac{C_2 q x}{\lambda x - \beta} \tag{7}$$

We use the Environmental Unified Identification Code System (EUIC[1]) of Taiwan Environmental Protection Administration (TEPA) as a pragmatic example to illustrate the computation of total cost for enhancing the data quality. The EUIC is an integrated data warehouse system that provides a single point of access to data extracted from four major TEPA databases, namely the Air Pollution Control System,

[1]http://euic.epa.gov.tw

q =0.25, represents that around one-fourth of data items might have quality problems.
λ =0.6, the average rate of improvements.
β =0.8, the degradation rate of data quality.
t_1 =4 months, defined by system managers, and shows that the data in the warehouse should be synchronized with its sources once every four months.
C_1 =1, coefficient for the ratio cost and data quality.
C_2 =\$3,000, the input man-hour cost.
C_3 =\$500, the cost of material resources such as computer hardware.

$$TC = \frac{1}{2}0.25(4) + 500x + \frac{0.25(4)^2}{2(0.6x - 0.8)} + \frac{3000(0.25)x}{0.6x - 0.8}$$

$$= \frac{1}{2} + \frac{300x^2 + 350x + 2}{0.6x - 0.8} \text{ where } x > \frac{4}{3}$$

Fig. 4 A simplified example for the computation of total cost

the Water Permit Database, the Hazardous Waste Control System, and the Toxic Release Database. According to the EUIC experience, maintaining the data in EUIC consistent with the legacy databases is the biggest problem in terms of data quality. Due to limited budget and resources, the EUIC managers usually face difficulties in determining how often to synchronize with legacy databases, and how many data items should be synchronized. We believe that Eq. (7) may assist the EUIC managers to evaluate the total cost for obtaining "fit to use" data quality. Fig. 4 describes a simplified computation of total cost using Eq. (7).

In order to determine how many poor-quality data items should be improved, given limited resources, we need to reduce the total cost. Let $\frac{dC}{dx}=0$, we can compute minimum total cost using the first order derivative of C.

$$x = \sqrt{\frac{C_1\lambda q^2 + 2C_2\beta q}{2C_3\lambda^2}} + \frac{\beta}{\lambda} \tag{8}$$

The x in Eq. (8) plays a fundamental role in the result of the cost/benefit modeling. It represents the amount of quality indicators that we may handle with the minimum cost. We observe that x consists of two parts where $\frac{\beta}{\lambda}$ represents the cost for improving inadequate data, i.e. β is the degradation rate of data quality and λ is the average improvement rate of each data quality indicator. It is obvious that the slope $\beta-\lambda x$ will be negative and has a chance crossing with axis t in Fig. 3 only if $x>\frac{\beta}{\lambda}$. Another factor for determining the amount of quality indicators is related to each parameter when we adopt the model. When the average improvement rate λ and the coefficient for improvement cost C_3 are increased, the amount of improved quality indicators decrease. Moreover, when the degradation rate of data quality is β, the degradation condition of data quality as the improvement started at q, and coefficient for lost cost C_1 are increased, the amount of improved quality indicators increase as well. There are some de facto factors to be considered as we adopt the model to evaluate the improvement plan for data quality. In general, C_1, C_2, C_3 can be viewed as constants, β and q can be obtained by examination, and λ can be decided by experienced data warehouse managers. We may gradually rectify the parameters in Eq. 8 for adapting the model to get close to de facto distributions.

V. RELATED WORK

Our work combines and enlarges the results from research in the field of data quality, data warehouses, and data integration. We mention here some relevant work and approaches. Before data warehouses drew the attention of researchers, the integration of heterogeneous data sources was investigated using semantic data models. For example, the TSIMMIS project (Chawathe *et al.*, 1994) has the goal of providing tools for integrated access to multiple and diverse information sources. Wrappers are used to encapsulate sources and repositories and mediators are used to find out the sources, which are suitably integrated and processed. Similarly, but with slightly different design strategies, InfoHub (Chu *et al.*, 1997) provides each data source a dedicated wrapper, which makes the overall architecture extensible and flexible. Furthermore, by pre-defined domain knowledge in the knowledge server, the mediator can plan to *pre-fetch* relevant information. This feature not only makes the system perform active services, but also improves the overall efficiency of the system. These integration systems are mainly focused upon improving the consistency of the global schema. Yet, a data warehouse deals with the problem in a broader way that, interestingly, makes things easier.

In data quality research, a number of studies have been done on the quality issues of information systems and data warehouse environments (DWQ, Jarke *et al.*, 1999; Kaplan *et al.*, 1998; Wang *et al.*, 1995a; Wang *et al.*, 1995b). Wang *et al.*, (1995b) proposed a framework of data quality analysis, based on the ISO 9000 standard. This framework reviews a significant part of the literature on data quality. Several studies have investigated the attribute-based model for data quality management (Wang *et al.*, 1995a; Rothenberg, 1996). Those approaches mainly augment databases with a *quality field* to form a new schema and perform organizational control over the processes that generate and modify data. However,

the data stored in *quality fields* is hard to separate for accessing or evaluating since it is tightly coupled with the other data fields in the same table. Hence, the query languages for databases such as SQL may need to be modified to fit the new schema architecture.

Kaplan *et al.* (1998) presented a decision support system to assist users to carry out data quality assessments of accounting information systems. Combining human judgment and the appropriate use of model-based algorithmic procedures, the system enabled users to decide the extent of testing and to select the minimum set of control procedures needed to ensure data reliability. Jarke *et al.* (1999) presented an approach to explore the architecture and the quality in data warehouses based upon extended repository. This approach extended the Goal-Question-Metric approach from software engineering to capture some quality dimensions in data warehousing architectures. The quality issues discussed in (Jarke *et al.* 1999) focused on the quality of the *design and implementation of data warehouses*, rather than the quality of the data stored in data warehouses.

VI. CONCLUSIONS

We have described an attribute-based metadata model, which separates the data in data warehouses into two aspects, "attribute data" and "quality data". We also explore how quality data can be represented as metadata and how it can be accessed in data warehousing architectures. Thus, data quality issues can be effectively managed and assured. Concerning the degradation of data quality in warehouses, we propose a cost/benefit model to perform the evaluation and find out what kinds of data items should be modified or improved. The main points in this paper can be summarized as follows.

1. A four-phase process is introduced for data quality management during the life cycle of data warehouses. As time goes on, data warehouse users may face the problem of interpretability, because they do not know how the data are transferred into the warehouse. Our approach allows system developers to document related quality data as metadata, which may be associated with the life cycle of data warehouses.
2. Quality requirements can be formally and clearly defined in terms of attribute-based metadata which provides diagnostic information to figure out the sources of data error.
3. Overall data quality conditions can be identified and relevant information can be provided for determining whether the data meet "fit to use" criteria and whether they need to be improved.
4. Users may filter the data retrieved from the warehouse based upon various quality requirements. On the basis of constraints of cost and timing, we may then figure out what kinds of data should be preferentially modified or improved in order to achieve maximum benefit of data quality.

We believe that data quality will be a significant issue as data warehousing becomes more and more popular. There is obviously a great deal of work to be done to obtain better data quality to support decision-making. One direction of current work will be continuing to expand the cost/benefit model for more detailed evaluation and analysis of data quality. For example, we may further analyze $\lambda(q)$ to explore the relationship between λ and the degree of data degradation q. In addition, applying AI techniques for data quality definition, and Machine Learning to enhance poor-quality data detection capability and identification may also be taken into account for future work.

ACKNOWLEDGEMENTS

This work was supported in part by National Science Council, R.O.C., grant# NSC88-2213-E011-004 and Taiwan Environmental Protection Administration, grant# EPA-88-U1L1-03-003. We would like to acknowledge the generosity of these organizations.

NOMENCLATURE

C	cost of time-independent issues
C_1, C_2, C_3	coefficients of cost
$\boldsymbol{IC}$	improvement cost
$\boldsymbol{LC}$	lost cost
q	data items that might have quality problems
$Q(t)$	degradation of data quality at time point t
$\boldsymbol{TC}$	total cost
W	amount of attribute data
x	amount of quality indicators that may handle with minimum cost

Greek Symbols

α	average rate of the improvement
β	degradation rate of data quality
Δ	data belonging to the attribute which can be associated with its quality metadata
λ	average rate of improvements
ρ	number of poor-quality data

REFERENCES

1. DWQ Project (2000) Available at http://www.dbnet.ece.ntua.gr/~dwq/.

2. Ballou, D. P., and Pazer, H. L., 1987, "Cost/Quality Tradeoffs for Control Procedures in Information Systems." *International Journal of Management Science*, Vol. 15, No. 6, pp. 509-521.
3. Barquin, R., and Edelstein, H., 1997, *Building, using, and managing the data warehouse*. PTR Publishing, New Jersey USA.
4. Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K, Papakonstantinou, Y., Ullman, J., and Widom, J., 1994, "The TSIMMIS Project: Integration of Heterogeneous Information Sources," *Proceedings of IPSI Conference*, Tokyo, Japan, pp. 7-18.
5. Chu, Y. C., Lien, C. C., and Yang, C. C., 1997, "InfoHub: A Flexible System for Retrieving and Integrating Heterogeneous Information Sources," *Proceedings of Workshop on Distributed System Technologies and Applications*, Taiwan, ROC, pp. 597-602.
6. Garcia-Molina, H., Labio, W. J., Wiener, J. L., and Zhuge, Y., 1999, "Distributed and Parallel Computing Issues in Data Warehousing," *In Proceedings of ACM Principles of Distributed Computing Conference*, pp. 7-10.
7. Huh, Y. U. Keller, F.R., Redman, T.C., and Watkins, A.R., 1990, "Data Quality," *Information and Software Technology*, Vol. 32, No. 8, pp. 559-565.
8. Jarke, M., Jeusfeld, M. A., Quix, C., and Vassiliadis, P, 1999, "Architecture and Quality in Data Warehouse: An Extended Repository Approach," *Information Systems*, Vol. 24, No. 3, pp.229-253.
9. Kimball, R., 1996, *The data warehouse toolkit*, Wiley Publishing, New York, USA.
10. Kaplan D., Krishnan, R., Padman, R., and Peters, J., 1998, "Assessing Data Quality in Accounting Information Systems," *Communications of the ACM*, Vol. 41, No. 2, pp. 72-78.
11. Orr K., 1998, "Data Quality and Systems Theory," *Communications of the ACM*, Vol. 41, No. 2, pp. 66-71.
12. Parsaye, K., and Chignell, M., 1993, *Intelligent database tools and applications: hyperinformation access, data quality, visualization, automatic discovery*, Wiley Publishing, New York, USA.
13. Rothenberg, J., 1996, "Metadata to Support Data Quality and Longevity," *Proceedings of First IEEE Metadata Conference*, Available at http://www.computer.org/conferences/meta96/rothenberg_paper/ieee.data-quality.html
14. Wand, Y., and Wang, R., 1996, "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, Vol. 39, No. 11, pp. 86-95.
15. Wang, R. Y., Reddy, M. P., and Kon, H. B., 1995a, "Toward Quality Data: an Attribute-based Approach," *Decision Support Systems*, Vol. 13, pp. 349-372.
16. Wang, R. Y., Storey, V. C., and Firth, C. P., 1995b, "A Framework for Analysis of Data Quality Research," *IEEE Trans. Knowledge and Data Engineering*, Vol. 7, No. 4, pp.623-640.
17. Yang, S. S., 1999, *An Evaluation Mechanism for Data Quality Assurance in Data Warehouse Environments*, Master Thesis, National Taiwan University of Science and Technology. (in Chinese).

Manuscript Received: Mar. 16, 2000
Revision Received: Nov. 23, 2000
and Accepted: Jan. 10, 2001

結合屬性詮釋及成本評估以強化資料倉儲環境中之資料品質

朱雨其[1] 楊珊珊[2] 楊鍵樵[1]

[1]國立台灣科技大學電子工程學系
[2]中華電信研究所

摘 要

資料品質之良窳，攸關資料倉儲系統發展的成敗。本文旨在探討資料倉儲環境中，如何評估系統內部所儲存資料之品質的方法，以及在提昇資料品質與降低成本支出間找尋合理之平衡點，期能以最低的成本花費，提供最優良的資料品質。我們倡議以屬性爲基礎的詮釋資料來確認與解釋資料品質，並發展一個四階段的資料品質管理作業流程，將其與資料倉儲系統的發展過程相溶合，尋求資料品質“最適使用”的狀態。其次，經由成本/效益評估模型的運算，可在有限成本控制下針對最嚴重的不良資料作優先的修正處理。本文倡議的方法不僅能加強屬性資料的詮釋性，免除資料誤用及語意的混淆，還能藉由成本/效益評估模型，爲資料品質的量化分析提供初步的作業基礎。

關鍵詞：資料倉儲，資料品質，詮釋資料，成本/效益評估模式。