# 4

# Modeling Information Quality Risk for Data Mining and Case Studies

Ying Su
*Information Quality Lab Resource Sharing Promotion Centre, Institute of Scientific and Technical Information of China, Beijing, China*

## 1. Introduction

Today, information is a vital business asset. For institutional and individual processes that depend on information, the quality of information (IQ) is one of the key determinants of the quality of their decisions and actions (Hand, et al., 2001; W. Kim et al., 2003; Mucksch, et al., 1996). Data mining (DM) technology can discover hidden relationships, patterns and interdependencies and generate rules to predict the correlations in data warehouses (Y. Su, et al., 2009c).

However, only a few companies have implemented these technologies because of their inability to clearly measure the quality of data and consequently the quality risk of information derived from the data warehouse(Fisher, et al., 2003). Without this ability it becomes difficult for companies to estimate the cost of poor information to the organization (D. Ballou, Madnick, & Wang, 2003). For the above reasons, the risk management of the IQ for DM is been identified as a critical issue for companies. Therefore, we develop a methodology to model the quality risk of information based on the quality of the source databases and associated DM processes.

The rest of this chapter is organized as follows. After a review of the relevant in Section 2, we introduce a forma1 model proposed for data warehousing and DM that attempts to support quality risks of different levels in Section 3. In section 4, we discuss the different quality risks that need to be considered for the output of Restriction operator, Projection and Cubic product operators. Section 5 describes an information quality assurance exercise undertaken for a finance company as part of a larger project in auto finance marketing. A methodology to estimate the effects of data accuracy, completeness and consistency on the data aggregate functions Count, Sum and Average is presented(Y. Su, et al., 2009a). The methodology should be of specific interest to quality assurance practitioners for projects that harvest warehouse data for decision support to the management. The assessment comprised ten checks in three broad categories, to ensure the quality of information collected over 1103 attributes. The assessment discovered four critical gaps in the data that had to be corrected before the data could be transitioned to the analysis phase. Section 6 applies above methodology to evaluate two information quality characteristics - accuracy and completeness - for the HIS database. Four quantitative measures are introduced to assess the risk of medical information quality. The methodology is illustrated through a medical domain: infection control. The results show the methodology was effective to detection and aversion of risk factors(Y. Su, et al., 2009b).

## 2. Literature review

### 2.1 IQ dimensions

Huang et al. (1999, p. 33) state that information quality has been conventionally described as how accurate information is. In the last couple of years, however, it has become clear that information quality encompasses multiple dimensions beyond accuracy. These dimensions can be gathered in various ways (Huang, et al., 1999). Huang et al. (1999) distinguish between three different approaches: the intuitive, systematic, and empirical one. The intuitive approach is one where IQ-criteria are based on the intuitive understanding or experience of one or several individuals. The main disadvantage of this approach is that it does not yield representative results. The systematic approach, according to Huang et al., focuses on how information may become deficient during the information production process. Few research strategies have followed this deductive-analytic or ontological approach (where real-life states are compared to the represented data states). One reason may be the fact that it is difficult to convey the results to information consumers. The third approach is an empirical one. Here, the criteria are gathered by asking large sets of information consumers about their understanding of information quality in specific contexts (as we have done with the online focus groups described earlier). The disadvantage of this approach, according to Huang et al. (1999, p. 34) is that the correctness or completeness of the results cannot be proven based on fundamental principles (as in the deductive systematic approach). There is also a risk, in Eppler's view, that the empirical results will not always be consistent or free of redundancies. It is also unclear, whether information consumers are always capable of articulating the information quality attributes which are important to them. Besides distinguishing the ways in which the criteria can be gathered, one can also distinguish the types of criteria that exist (Eppler, 2006).

The coexistence of these different criteria to IQ in business processes may result in conflicting views of IQ among information providers and consumers. These differences can cause serous breakdowns in communications both among information suppliers and between information suppliers and consumers. But even with improved communication among them, each of the principal approaches to IQ shares a common problem: each offers only a partial and sometimes vague view of the basic elements of IQ.

In order to fully exploit favourable conditions of these criteria and avoid unfavourable ones, we present a definition approach of IQ that is based on characteristics of enterprise activities precedence relationship between them (Table 1.). Enterprise activities are processing steps within a process transforming objects and requiring resources for their execution. An activity can be classified as a structured activity if it is computable and controllable. Otherwise, it is categorized as a non-structured activity. Accounting, planning, inventory control, and scheduling activities are examples of structured activities. Typical examples of non-structured activities are human-based activities such as design, reasoning, or thinking activities.Table 1. gives the reference dimensions of upstream activity regarding the context in the business processes (Su & Jin, 2006).

Su and Jin summarized academic research on the multiple dimensions of IQ, and assigned the four cases based on types of relationship of enterprise activities, as the second and third columns of Table 1. The fifth column of Table 1. summarizes academic research on the multiple dimensions of IQ. The first row is Ballou and Pazer's (1985) study, which takes an empirical, market research approach of collecting data from information consumers to determine the dimensions of importance to them. Table 1. lists the dimensions uncovered in Zmud's (1978) pioneering IQ research study, which considers the dimensions of information

important to users of hard-copy reports. Because of the focus on reports, information accessibility dimensions, which are critical with on-line information, were not relevant.

| Activity Taxonomy | Upstream Activity | Downstream Activity | Definition Approach | Reference Dimensions of IQ for Upstream Activity |
|---|---|---|---|---|
| CASE Ⅰ | Non-Structured | Non-Structured | User-based | Consistent representation, Interpretability, Case of understanding, Concise representation, Timeliness, Completeness (Ballou & Pazer, 1985), Value-added, relevance, appropriate, Meaningfulness, Lack of confusion (Goodhue, 1995). Arrangement, Readable, Reasonable (Zmud, 1978). |
| CASE Ⅱ | Non-Structured | Structured | Intuitive | Precision, Reliability, freedom from bias (DeLone & McLean, 2003). |
| CASE Ⅲ | Structured | Non-Structured | User-based | See also CASE Ⅰ |
| CASE Ⅳ | Structured | Structured | System | Data Deficiency, Design Deficiencies, Operation Deficiencies (Huang et al., 1999). |
| Inherent IQ | Accuracy, Cost, Objectivity, Believability, Reputation, Accessibility, Correctness (Wang & Strong, 1996), Unambiguous (Wand & Wang, 1996). Consistency (English, 1999). | | | |

Table 1. Activity-based defining to the IQ dimensions

In our analysis, we consider risks associated with two well-documented information quality attributes: accuracy and completeness. Accuracy is defined as conformity with the real world. Completeness is defined as availability of all relevant data to satisfy the user requirement. Although many other information quality attributes have been introduced and discussed in the existing literature, these two are the most widely cited. Furthermore, accuracy and completeness can be measured in an objective manner, something that is usually not possible for other quality attributes.

### 2.2 Overview of BDM and data warehousing

Business data mining (BDM), also known as "knowledge discovery in databases"(Bose & Mahapatra, 2001), is the process of discovering interesting patterns in databases that are useful in decision making. Business data mining is a discipline of growing interest and importance, and an application area that can provide significant competitive advantage to an organization by exploiting the potential of large data warehouses.

In the past decade, BDM has changed the discipline of information science, which investigates the properties of information and the methods and techniques used in the acquisition, analysis, organization, dissemination and use of information (Chen & Liu, 2004).

BDM can be used to carry out many types of task. Based on the types of knowledge to be discovered, it can be broadly divided into supervised discovery and unsupervised discovery. The former requires the data to be pre-classified. Each item is associated with a unique label, signifying the class in which the item belongs. In contrast, the latter does not require pre-classification of the data and can form groups that share common characteristics. To carry out these two main task types, four business data mining approaches are commonly used: clustering(Shao & Krishnamurty, 2008), classification(Mohamadi, et al., 2008), association rules(Mitra & Chaudhuri, 2006) and visualization (Compieta et. al., 2007).
As mentioned above, BDM can be used to carry out various types of tasks, using approaches such as classification, clustering, association rules, and visualization. These tasks have been implemented in many application domains. The main application domains that BDM can support in the field of information science include personalized environments, electronic commerce, and search engines. Table 2. summarizes the main contributions of BDM in each application.
A data warehouse can be defined as a repository of historical data used to support decision making (Sen & Sinha, 2007). BDM refers to the technology that allows the user to efficiently retrieve information from the data warehouse (Sen, et al., 2006).
The multidimensional data model or data cube is a popular model used to conceptualize the data in a data warehouse (Jin, et al., 2005). We emphasize that the data cube that we are referring to here is a data model, and is not to be confused with the well-known CUBE operator, which performs extended grouping and aggregation.

| Application | Approaches | Contributions |
|---|---|---|
| Personalized Environments | Usage mining | To adapt content presentation and navigation support based on each individual's characteristics. |
| | Usage mining with collaborative filtering | To understand users' access patterns by mining the data collected from log files. |
| | Usage mining with content mining | To tailor to the users' perceived preferences by matching usage and content profiles. |
| Electronic Commerce | Customer management | To divide the customers into several segments based on their similar purchasing behavior. |
| | Retail business | To explore the association structure between the sales of different products. |
| | Time series analysis | To discover patterns and predict future values by analyzing time series data. |
| Search Engine | Ranking of pages | To identify the ranking of the pages by analyzing the interconnections of a series of related pages. |
| | Improvement of precision | To improve the precision by examining textual content and user's logs. |
| | Citation analyses | To recognize the intellectual structure of works by analyzing how authors are cited together. |

Table 2. Business data mining contributions

### 2.3 Research contributions

The main contribution of this research is the development of a rigorous methodology to confirm the information quality risks of data warehouses. Although little formal analysis of this nature has been addressed in previous research, two approaches proposed earlier have influenced our work. Michalski, G. (2008) provides a methodology to determine the level of accounts receivable using the portfolio management theory in a firm. He presents the consequences that can result from operating risk that is related to purchasers using payment postponement for goods and/or services, however, he don't provide a methodology for deriving quality risks for the BDM (Michalski, 2008). Cowell, R. G., Verrall, R. J., & Yoon, Y. K. (2007) construct a Bayesian network that models various risk factors and their combination into an overall loss distribution. Using this model, they show how established Bayesian network methodology can be applied to: (1) form posterior marginal distributions of variables based on evidence, (2) simulate scenarios, (3) update the parameters of the model using data, and (4) quantify in real-time how well the model predictions compare to actual data (Cowell, et al., 2007).

## 3. The cube model and risks

### 3.1 Basic definitions

A data cube is the fundamental underlying construct of the multidimensional database and serves as the basic unit of input and output for all operators defined on a multidimensional database. It is defined as a 6-tuple, $\langle C, A, f, d, O, L \rangle$ where the six components indicate the characteristics of the cube. These characteristics are:

- $C$ is a set of $m$ characteristics $C = \{c_1, c_2, \cdots c_m\}$ where each $c_i$ is a characteristic having domain (dom) C.;
- $A$ is a set of t attributes $A = \{a_1, a_2, \cdots a_t\}$ where each $a_i$ is an attribute name having domain Dom A. We assume that there exists an arbitrary total order on A, $\leq A$. Thus, the attributes in A (and any subset of A) can be listed according to $\leq A$. Moreover we say that each $a_i \in A$ is recognizable to the cube C;
- $f$ is a one-to-one mapping, $f : C \to 2^A$, which maps a set of attributes to each characteristic. a set of attributes to each characteristic. The mapping is such that attribute sets corresponding to characteristics are pairwise disjoint, i.e., $\forall i, j, i \neq j, f(c_i) \bigcap f(c_j) = \varnothing$. Also, all attributes are mapped to characteristics (i.e., $\forall x, x \in A, \exists c, c \in C, x \in f(c)$). Hence, $f$ partitions the set of attributes among the characteristics. We refer to f(c) as the schema of c;
- d is a Boolean-valued function that partitions C into a set of dimensions D and a set of dimensions D and a set of measures M. Thus, $C = D \bigcup M$ where $D \bigcap M = \varnothing$. The function d is *defined is follows*:; $\forall x \in C, d(x) = \begin{cases} 1 & if \quad x \in D \\ 0 & otherwise \end{cases}$
- $O$ is a set of partial orders such that each $o_i \in O$ is a partial order defined on $f(c_i)$ and $|O| = |C|$.
- $L$ is a set of cube cells. A cube cell is represented as an $\langle address, content \rangle$ pair. The address in this pair is an n-tuple, $\langle \alpha_1, \alpha_2, \cdots \alpha_n \rangle$, where n is the number of dimensional attributes in the cube, i.e., $n = |A_d|$. The content of a cube cell is defined similarly. It is a k-tuple, $\langle \chi_1, \chi_2, \cdots \chi_n \rangle$, where k is the number of metric attributes in the cube; i.e.,

$k = |A_m|$, where $A_m$, represents the set of all metric attributes; For notational convenience, we denote the structural address component of L as *L.AC* and the structural content component as *L.CC*. We denote the ith address value component of cube cell *1* as *1.AC[i]* and the *i*th content value component as *l.CC[i]*.

We now provide an example to clarify this definition. Subsequently, this will be used as a running example for the rest of the chapter. Consider a cube Sales which represents a multidimensional database of sales figures of certain products. The Sales cube has the following features (note the correspondence of the example to the definition above).

- The data are described by the characteristics time, product, location, and sales. Hence, the cube has a characteristics set C = $\{$product,time, address,sales$\}$ (m=4).
- The time characteristic is described by the attributes day, week, month, and year; the product characteristic is described by the product_id, weight and name attributes; the location characteristic is described by the store_name, store_address, state, and region attributes. The sales characteristic is described by the store_sales and store_cost attributes. Thus, for the Sales cube, A= { day, week, month, year, product_id, weight, name, store_name, store_address, state, region, store_sales, store_cost } (t = 13).
- Each of the characteristics, as explained in the previous item, are described by specific attributes. In other words, for the Sales cube, the mapping f is as follows:

$f(\text{time}) = \{ \text{day, week, month, year} \}$

$f(\text{product}) = \{ \text{product\_id, weight, name} \}$

$f(\text{location}) = \{ \text{store\_name, store\_address, state, region} \}$

$f(\text{sales}) = \{ \text{store\_sales, store\_cost} \}$

Also note that the attribute sets shown above re mutually disjoint.

· An example of a partial order in *O* on the Sales given by the following:

$O_{\text{time}} = \{ \langle \text{day, week} \rangle, \langle \text{day, month} \rangle, \langle \text{day, year} \rangle, \langle \text{month, year} \rangle \}$

$O_{\text{product}} = \{ \langle \text{product\_id, name} \rangle, \langle \text{product\_id, weight} \rangle \}$

$O_{\text{location}} = \{ \text{store\_name, store\_address, state, region} \}$

$O_{\text{sales}} = \{ \}$

· To present a simple example of L, we assume the following attributes and corresponding domains for the Sales cube data:

A= { year, product_id, store_address, store_sales, store_cost }

Dom year= { 2001,2002,2003,2004 }

Dom product_id= { P1, P2, P3, P4 }

Dom store_address = { "Valley View","Valley Ave",Coit Rd.", "Indigo Ct" }

Dom store_sales $\in R$

Dom store_cost $\in R$

· Then an element $l \in L$ may be expressed as follows:

$$l = \langle l.AC, l.CC \rangle \text{ where:}$$

$$l.AC = \langle 2001, P1, \text{"4 Valley View"} \rangle \text{, corresponding to the structural components:}$$

$$/.\text{AC} = \langle \text{year, product\_id, store\_address} \rangle$$

$$l.CC = \langle 30, 120 \rangle \text{, corresponding to the structural components:}$$

$$l.CC = \langle \text{store\_sales, store\_cost} \rangle$$

A possible cube using the data from above is shown below pictorially in Fig 1. Henceforth, we will work with cubes in the development of theory in this chapter.

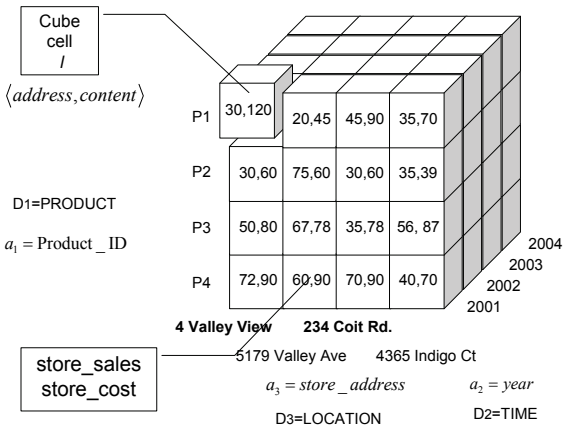



Fig. 1. Data Cube Example with Notation

Consider a cube *C* that contains tuples captured for a predefined real world entity type. Each tuple in *C* is either accurate, inaccurate, nonmember, or an incomplete. These terms are formally defined below:

- A tuple is accurate if all of its attribute values are accurate.
- A tuple is inaccurate if it has one or more inaccurate (or null) values for its nonidentifier attributes, and no inaccurate values for its identifier attribute(s).
- A tuple is a nonmember if it should not have been captured into C but it is. A nonmembership tuple might have inaccurate values either in its identifier attributes or nonidentifier ones which is mistakenly included in the cube.
- A tuple belongs to the incomplete set if it should have been captured into C but it is not.

We denote the set of accurate, inaccurate, nonmember and incomplete tuples by $C_A$, $C_I$, $C_N$, and $C_C$ respectively. Then, we use the notion of a conceptual cube T in order to understand the relationship between tuples in C and the underlying entity instances in the real world. Cube T consists of tuples as they should have been captured in C if there were no errors in an ideal world. Tuples in T belong to three categories as follows:

- $T_A$, the set of instances in T that are correctly captured into C and thus remain accurate;
- $T_I$, the set of instances in T that are captured into C, and one or more of their nonidentifying attribute values are inaccurate or null;

- $T_C$, the set of instances in T that have not been captured into C and therefore form the incomplete dataset for C.

### 3.2 Cube-level risks

Based on the above definitions, we define the following quality risks for a cube $C$. $|L|$, $|L_A|$, $|L_I|$ $|L_N|$, and $|L_C|$ denote the cardinalities of the sets $L$, $L_A$, $L_I$, $L_N$, and $L_C$, respectively.

- Accuracy of $C$, measured as $\Pr_A(C) = |L_A|/|L|$, is the probability that a tuple in $L$ accurately represents an entity in the real world.
- Inaccuracy of $C$, measured as $\Pr_I(C) = |L_I|/|L|$, is the probability that a tuple in $L$ is inaccurate.
- Nonmembership of $C$, measured as $\Pr_N(C) = |L_N|/|L|$, is the probability that a tuple in $C$ is a nonmember.
- Incompleteness of $C$, measured as $\Pr_C(C) = |L_C|/(|L| - |L_N| + |L_C|)$, is the probability that an information resource in the real world is not captured in $C$.

The data cube is a data model for representing business information using multidimensional database (MDDB) technology. The following example about a cube Sale illustrates these risks. Table 3. hows the data stored in the feature class $C$, and Table 4. shows the incomplete information for $C$. The attribute set $\{Time\_ID, Customer\_ID, Store\_Address\}$

| Product_ID | Time_ID | Customer_ID | Store_Address | Store_Cost | Store_Sales | Tuple Status |
|---|---|---|---|---|---|---|
| 1 | 2001 | 334-1626-003 | 5203 Catanzaro Way | 10,031 | 100 | A |
| 2 | 2003 | 334-1626-001 | 1501 Ramsey Circle | 7,342 | 200 | A |
| 3 | 2002 | 334-1626-004 | 433 St George Dr | **9,254** | 300 | I |
| 4 | 2004 | 334-1626-005 | 1250 Coggins Drive | 8,856 | 250 | A |
| 5 | 2000 | 334-1626-006 | **4 Valley View** | 8,277 | 120 | I |
| 6 | 1999 | 334-1626-007 | 5179 Valley Ave | 9,975 | 360 | A |
| **7** | **2002** | **334-1626-012** | **234 Coit Rd.** | **8,230** | **640** | **N** |
| 8 | **2004** | 334-1626-002 | 4365 Indigo Ct | 1,450 | 210 | I |
| 9 | 2005 | **334-1626-019** | 5006 Highland Drive | 8,645 | 780 | I |

Table 3. Feature Class Cube C

| ID | Time_ID | Customer_ID | Store_Address | Store_Cost | Store_Sales | Tuple Status |
|---|---|---|---|---|---|---|
| 10 | 2004 | 334-1626-008 | 321 herry Ct. | 11,412 | 365 | C |

Table 4. Incomplete Cube $L_C$

| ID | Rows Status | Error Description |
|---|---|---|
| 3 | Inaccurate | Store_Cost should be "9,031" |
| 5 | Inaccurate | Store_Address should be "6 Valley View" |
| 7 | Nonmember | Should not belong to cube C |
| 8 | Inaccurate | Store_Sales should be "790" |
| 9 | Inaccurate | Customer_ID should be "334-1626-009" |

Table 5. Errors Cube in L

| Cube | Size | $\Pr_A(C)$ | $\Pr_I(C)$ | $\Pr_N(C)$ | $\Pr_C(C)$ |
|---|---|---|---|---|---|
| C | 9 | 0.44 | 0.44 | 0.11 | 0.11 |

Table 6. Quality Profile for Cube C

forms the address for *C*. The Tuple Status column in Table 3. indicates whether a tuple is accurate (A), inaccurate (I), or a nonmember (N). Cells in *C* that are set in bold type contain inaccurate values, and the row set in bold type is a nonmember. Table 5. describes errors in *C*, and Table 6. provides the quality measures.

### 3.3 Risk measures for attribute-level

To assess the quality metrics of derived cubes based on the quality profile for the input cube, we need to estimate quality metrics at the attribute level for some of the relational operations. Let $K_C$ and $Q_C$ be the set of identifier and nonidentifier attributes of C. Furthermore, let $k_C$ and $q_C$ be the number of identifier and nonidentifier attributes, respectively. We make the following assumptions regarding the quality metrics for attributes of C.

Assumption 1. Error probabilities for identifier (nonidentifier) attributes are identically distributed. Error probabilities for all attributes are independent of each other.

Assumption 2. The probability of an error occurring in a nonidentifier attribute of a nonmember tuple is the same as the probability of such an error in any other tuple.

Let $\Pr_A(K_C)$ denote accuracy for the set of attributes $K_C$, and $\Pr_{Aa}(K_C)$ denote accuracy of each attribute in $K_S$. Thus $\Pr_A(K_C) = \left(|L_A| + |L_I|\right) \big/ |L| = \Pr_A + \Pr_I$. From Assumption 1, we have $\Pr_A(K_C) = \Pr_A(k)^{k_C}$, and, therefore

$$\Pr_{Aa}(K_C) = (\Pr_A(C) + \Pr_I(C))^{1/k_S} \tag{1}$$

Let $\alpha_{Q_S}$ denote accuracy for the set of attributes $Q_C$ and $\Pr_{Aa}(Q_C)$ denote accuracy of each attribute in $Q_C$. From assumption 2, we have $\Pr_A(Q_C) = \dfrac{|L_A|}{|L_A| + |L_I|} = \dfrac{\Pr_A(C)}{\Pr_A(C) + \Pr_I(C)}$. Because there are $q_C$ nonidentifier attributes, we have $\Pr_A(Q_C) = \Pr_{Aa}(Q_C)^{q_C}$ and therefore

$$\Pr_{Aa}(Q_C) = \left(\frac{\Pr_A(C)}{\Pr_A(C) + \Pr_I(C)}\right)^{1/q_C} \tag{2}$$

## 4. Cube-level risks for proposed operations

### 4.1 Selection operation

The selection operator restricts the values on one or more attributes based on specified conditions, where a given condition is in the form of a predicate. Thus, a set of predicates is evaluated on selected attributes, and cube cells are retrieved only if they satisfy a given predicate. If there are no cube cells that satisfy P, the result is an empty cube. The algebra of the selection operator is then defined as follows:

Input: A cube $C_I = \langle C, A, f, d, O, L \rangle$ and a compound predicate P.

Output: A cube $C_O = \langle C, A, f, d, O, L_O \rangle$ where $L_0 \subseteq L$ and $L_0 = \{l | (l \in L) \wedge (l \text{ satisfies P})\}$
Mathematical Notation:

$$\sigma_P C_I = C_O \tag{3}$$

We define a conceptual cube (denoted by $U$) that is obtained by applying the predicate condition to the conceptual cube $T$. $U_j$ denotes instances in $T_j$ that satisfy the predicate condition for j = A, I, and C. Fig 1. shows the mapping between the subsets of the conceptual and stored and cubes. We make two assumptions that are widely applicable.

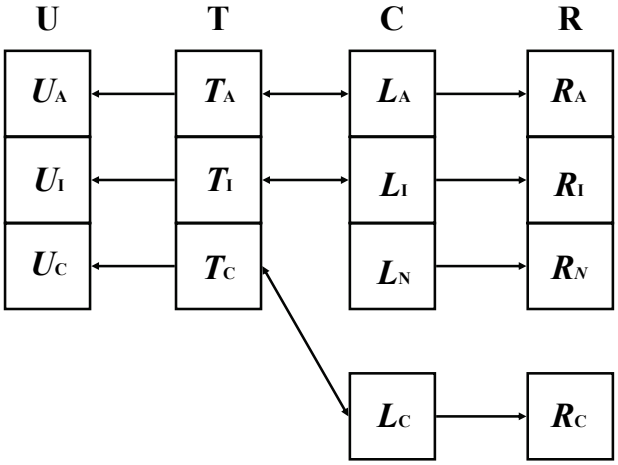



Fig. 2. Mapping Relations between the Concept and Physical

Assumption 3. Each true attribute value of an entity instance is a random (not necessarily uniformly distributed) realization from an appropriate underlying domain.
We then have

$$\frac{|U|}{|T|} = \frac{|U_A|}{|T_A|} = \frac{|U_I|}{|T_I|} = \frac{|U_C|}{|T_C|} \tag{4}$$

Assumption 4. The occurrences of errors in C are not systematic, or, if they are systematic, the cause of the errors is unknown.
This implies that the inaccurate attribute values stored in C are also random realizations of the underlying domains. It follows that

$$\frac{|U|}{|T|} = \frac{|R|}{|L|} = \frac{|R_{L_A}|}{|L_A|} = \frac{|R_{L_A}|}{|L_A|} = \frac{|R_{L_I}|}{|L_I|} = \frac{|R_{L_N}|}{|L_N|} = \frac{|R_{L_C}|}{|L_C|} \tag{5}$$

First, we consider the inequality condition. To illustrate this scenario, we use the cubes C and $L_C$ as shown in Table 3. and Table 4. Consider a query to retrieve tuples on feature class whose Customer_ID end with letters that evaluates to greater than "005". R and $R_C$ are shown in Table 7. and 2, respectively. $R_A$, $R_I$, and $R_N$ refer to accurate, inaccurate, and nonmember subsets of R.

After query execution, all accurate tuples satisfying the predicate condition remain accurate in *R*. Similarly, all selected inaccurate and nonmember tuples continue to be inaccurate and nonmember in R, respectively. Tuples belonging to the incomplete dataset $L_C$ that would have satisfied the predicate condition now become part of $R_C$, the incomplete set for R. Therefore, there is no change in the tuple status for the selected tuples. The expected value of $|R_{L_A}|$ is $|L_A|\bullet|R|/|L|$. Similarly, we have $|R_{L_I}| = |L_I|\bullet|R|/|L|$, $|R_{L_N}| = |L_N|\bullet|R|/|L|$, and $|R_{L_C}| = |L_C|\bullet|R|/|L|$. Using these identities in the definitions of $\Pr_A(R), \Pr_I(R), \Pr_N(R)$ and $\Pr_C(R)$, it is easily seen that $\Pr_A(R) = \Pr_A(C)$, $\Pr_I(R) = \Pr_I(C)$ $\Pr_N(R) = \Pr_N(C)$ and $\Pr_C(R) = \Pr_C(C)$. We show the algebra for $\Pr_C(R)$ here:

| Product_ID | Customer_ID | Store_Address. | Store_Cost | Store_Sales | Tuple Status |
|---|---|---|---|---|---|
| 4 | 334-1626-005 | 1250 Coggins Drive | 8,856 | 250 | A |
| 5 | 334-1626-006 | **4 Valley View** | 8,277 | 120 | I |
| 6 | 334-1626-007 | 5179 Valley Ave | 9,975 | 360 | A |
| 7 | **334-1626-012** | **234 Coit Rd.** | **8,230** | **640** | **N** |
| 8 | 334-1626-002 | 4365 Indigo Ct | 1,450 | 210 | I |
| 9 | **334-1626-019** | 5006 Highland Drive | 8,645 | 780 | I |

Table 7. Query Result R for Selection Operation

$$\Pr_C(R) = \frac{|R_C|}{\left(|R| - |R_N| + |R_C|\right)} = \frac{|R|\bullet|L_C|/|L|}{|R|\left[1 - \left(|L_N|/|L|\right) + \left(|L_C|/|L|\right)\right]} = \frac{|L_C|}{\left(|L| - |L_N| + |L_C|\right)} \tag{6}$$

### 4.2 Projection operation

The risky projection operator restricts the output of a cube to include only a subset of the original set of measures. Let S be a set of projection attributes such that $S \subseteq A_m$. Then the output of the resulting cube includes only those measures in C. The algebra of risky projection is defined as follows:

Input: A cube $C_I = \langle C, A, f, d, O, L\rangle$ and a set of projection attributes C.

Output: A cube $C_O = \langle C, A_O, f_O, d, O, L_O\rangle$ where $A_O = S \cup A_d$, $f_O : C \to 2^{A_O}$, such that $f_O(c) = f(c) \cup A_O$, and $L_O = \left\{l_O \middle| \exists l \in L, l_O.AC = l.AC, l_O.CC = \langle l.CC[s_1], l.CC[s_2], \cdots, l.CC[s_n]\rangle\right\}$, where $\{s_1, s_2, \cdots, s_n\} = S$.

Mathematical Notation:

$$\Pi_S C_I = C_O \tag{7}$$

Fig. 3 illustrates the mapping between tuples in *C* and *R*. The notation $L_{I\to A}$, $L_{I\to I}$, and $L_{I\to N}$ refer to those inaccurate tuples in *C* that become accurate, remain inaccurate, and become nonmembers, respectively, in R. Each tuple in $L_{I\to N}$ contributes a corresponding tuple to the incomplete dataset $R_C$; we denote this contribution by $L_{I\to C}$. We denote by $k_p$ and $q_p$ the number of address and content attributes of *C* that are projected into *R*.

We estimate the sizes of the various subsets of R and of the set $R_C$ using the attribute-level quality metrics derived in Equality (1) and (2). These sizes depend on the cardinality of the

identifier for the resulting cube, and whether or not these attributes were part of the identifier of the original cube. Let $k_R$ and $q_R$ denote the number of identifier and nonidentifier attributes of R. We further define the following:

- $k_{p\to k}$ : Number of projected identifier attributes of *C* that are part of the identifier for R.
- $q_{p\to k}$ : Number of projected nonidentifier attributes of *C* that become part of the identifier for R.

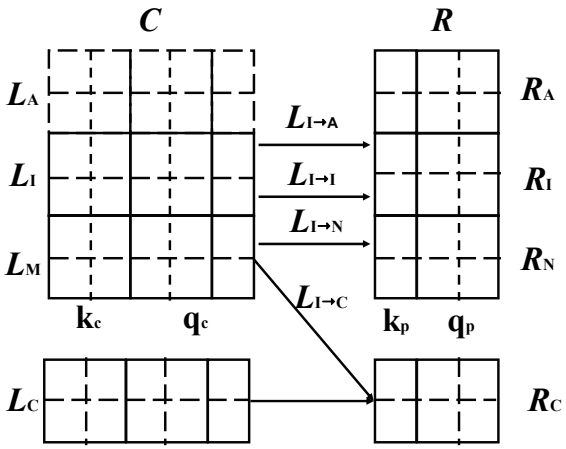



Fig. 3. Tuple Transformations for the Projection Operation

- $k_{p\to Q}$ : Number of projected identifier attributes of *C* that become part of nonidentifiers of R.
- $q_{p\to Q}$ : Number of projected nonidentifier attributes of *C* that are nonidentifier attributes of R.

The following equalities follow from our definitions:

$$k_p = k_{p\to k} + k_{p\to Q},\ q_p = q_{p\to k} + q_{p\to Q},$$

$$k_R = k_{p\to k} + q_{p\to K},\ and\ q_R = k_{p\to Q} + q_{p\to Q}.$$

A tuple in R is accurate only if all values of the projected attributes are accurate. From Equality (1), we know that each projected identifier attribute of C has accuracy $\Pr_{Aa}(K_C)$, whereas each projected nonidentifier attribute of C has an accuracy of $\Pr_{Aa}(Q_C)$ (2). The probability that a tuple is accurate in R is therefore given by

$$\begin{aligned}\Pr_A(R) &= \left[\left(\Pr_A(C)+\Pr_I(C)\right)^{1/k_C}\right]^{k_{p\to K}} \bullet \Pr_{Aa}(Q_C)^{q_{p\to K}} \bullet \left[\left(\Pr_A(C)+\Pr_I(C)\right)^{1/k_C}\right]^{k_{p\to Q}} \bullet \Pr_{Aa}(Q_C)^{q_{p\to Q}} \\ &= \left(\Pr_A(C)+\Pr_I(C)\right)^{\left(k_{p\to K}+k_{p\to Q}\right)/k_C} \bullet \Pr_{Aa}(Q_C)^{q_{p\to K}+q_{p\to Q}}\end{aligned} \quad (8)$$

Tuples in R are inaccurate if all the identifying attributes in R have accurate values, and at least one of the nonidentifying attributes of R is inaccurate. The size of the inaccurate set of

R can therefore be viewed as the difference between the set of tuples with accurate identifying attribute values and the set of accurate tuples. The former, which corresponds to $|R|(\mathrm{Pr}_A(R)+\mathrm{Pr}_I(R))$, is equal to $|R|(\mathrm{Pr}_A(C)+\mathrm{Pr}_I(C))^{k_{p\to K}/k_S} \bullet \mathrm{Pr}_{Aa}(Q_C)^{q_{p\to K}}$. It then follows that

$$\begin{aligned}\mathrm{Pr}_I(R) = &\left[(\mathrm{Pr}_A(C)+\mathrm{Pr}_I(C))^{k_{p\to K}/k_C} \bullet \mathrm{Pr}_{Aa}(Q_C)^{q_{p\to K}}\right] \\ &\bullet\left[1-(\mathrm{Pr}_A(C)+\mathrm{Pr}_I(C))^{k_{p\to Q}/k_C} \bullet \mathrm{Pr}_{Aa}(Q_C)^{q_{p\to Q}}\right]\end{aligned} \quad (9)$$

Using the equality $\mathrm{Pr}_N(R)=1-\mathrm{Pr}_A(R)-\mathrm{Pr}_I(R)$, the nonmembership for R is obtained as $\mathrm{Pr}_N(R)=1-(\mathrm{Pr}_A(C)+\mathrm{Pr}_I(C))^{k_{p\to K}/k_C} \bullet \mathrm{Pr}_{Aa}(Q_C)^{q_{p\to K}}$.

The incomplete dataset $R_C$ consists of the two parts: (i) tuples resulting from $L_C$ and (ii) the inaccurate tuples in C that become nonmembers in R and contribute to $R_C$. Because $|L_{I\to C}|=|L_{I\to N}|$, we determine $|L_{I\to C}|$ as $|R_N|-|L_N|$.Nothing that $|R|=|L|$, it follows that

$$\begin{aligned}|R_C| &= |L_C|+|R_N|-|L_N| = |L| \bullet \frac{\mathrm{Pr}_C(L)\bullet(1-\mathrm{Pr}_N(L))}{1-\mathrm{Pr}_C(L)} + (\mathrm{Pr}_N(R)\bullet|R|)-(\mathrm{Pr}_N(C)\bullet|L|) \\ &= |R|\left[\mathrm{Pr}_C(L)-\mathrm{Pr}_N(L)+\mathrm{Pr}_N(R)(1-\mathrm{Pr}_C(L))\right]/(1-\mathrm{Pr}_C(L))\end{aligned} \quad (10)$$

Substituting $|R_C|$ in the definition of (6), after some algebraic simplification, this yields $\mathrm{Pr}_C(R)=[\mathrm{Pr}_C(C)-\mathrm{Pr}_N(C)+\mathrm{Pr}_N(R)(1-\mathrm{Pr}_C(C))]/(1-\mathrm{Pr}_N(C))$.

Because $\mathrm{Pr}_N(R)=1-\mathrm{Pr}_A(R)-\mathrm{Pr}_I(R)$, we have

$$\begin{aligned}&\mathrm{Pr}_C(R)=[\mathrm{Pr}_C(C)-\mathrm{Pr}_N(C)+(1-\mathrm{Pr}_A(R)-\mathrm{Pr}_I(R))(1-\mathrm{Pr}_C(C))]/(1-\mathrm{Pr}_N(C)) \\ &=1-\frac{1-\mathrm{Pr}_C(C)}{1-\mu_C}(\mathrm{Pr}_A(R)-\mathrm{Pr}_I(R)) \\ &=1-\frac{1-\mathrm{Pr}_C(S)}{1-\mathrm{Pr}_N(C)}\bullet\left[(\mathrm{Pr}_A(C)+\mathrm{Pr}_I(C))^{k_{p\to K}/k_C} \bullet \mathrm{Pr}_{Aa}(Q_C)^{q_{p\to K}}\right]\end{aligned} \quad (11)$$

### 4.3 Cubic product operation

The Cubic Product operator is a binary operator that can be used to relate any two cubes. Often it is useful to combine the information in two cubes to answer certain queries (which we will illustrate with an example). The algebra of the Cubic Product operator is defined as follows:

Input: A cube $C_1=\langle C_1,A_1,\mathrm{f}_1,\mathrm{d}_1,O_1,L_1\rangle$ and a cube $C_2=\langle C_2,A_2,\mathrm{f}_2,\mathrm{d}_2,O_2,L_2\rangle$

Output: A cube $C_O=\langle C_0,A_0,\mathrm{f}_0,\mathrm{d}_0,O_0,L_0\rangle$, where $C_0=\Lambda_{C1}(C_1)\cup\Lambda_{C2}(C_2)$;
$A_0=\Lambda_{C1}(A_1)\cup\Lambda_{C2}(A_2)$;
$L_0=\{l_0|\exists l_1,\exists l_2,l_1\in L_1,l_2\in L_2,l_0.AC=l_1.AC\bullet l_2.AC,l_0.CC=l_1.CC\bullet l_2.CC\}$ where $l_1.AC\bullet l_2.AC$ denotes the concatenation of $l_1.AC$ and $l_2.AC$. In addition:
$\forall c_i\in(C_1\cup C_2)$

$$f_O = \begin{cases} f_1 & when\ \ applied\ \ to\ \ c_i \in C_1 \bullet c_i \\ f_2 & when\ \ applied\ \ to\ \ c_j \in C_2 \bullet c_i \end{cases} \qquad \forall c_i \in (C_1 \cup C_2)$$

$$d_O = \begin{cases} d_1 & when\ \ applied\ \ to\ \ c_i \in C_1 \bullet c_i \\ d_2 & when\ \ applied\ \ to\ \ c_j \in C_2 \bullet c_i \end{cases}$$

$$\forall a_i \in (f(C_1) \cup f(C_2))$$

$$O_O = \begin{cases} O_1 & when\ \ applied\ \ to\ \ a_i \in f(C_1) \\ O_2 & when\ \ applied\ \ to\ \ a_j \in f(C_2) \end{cases}$$

Mathematical Notation:

$$C_1 \otimes C_2 = C_O \tag{12}$$

To evaluate the quality profile for the Cartesian product R of two specified cubes (say $C_1$ and $C_2$), we first need a basis to categorize tuples in R as accurate, inaccurate, and nonmember, and to identify tuples that belong to the incomplete dataset of R. To illustrate this, Let Feature and Employee Table are the two realized cubes with tuples as shown in Table 8. and Table 9.

| Product_ID | Time_ID | Customer_ID | Store_Address | Store_Cost | Store_Sales | Status |
|---|---|---|---|---|---|---|
| P1 | 2001 | 334-1626-003 | 5203 Catanzaro Way | 10,031 | 100 | A |
| P2 | 2000 | 334-1626-006 | **4 Valley View** | 8,277 | 120 | I |
| **P3** | **2002** | **334-1626-012** | **234 Coit Rd.** | **8,230** | **640** | N |
| P5 | 2004 | 334-1626-008 | 321 herry Ct. | 11,412 | 365 | C |
| P4 | 2004 | 334-1626-005 | 1250 Coggins Drive | 8,856 | 250 | A |

Table 8. Actual Data Captured on Feature Table

| Employee_ID | Employee_Name | Position_Title | Tuple Status |
|---|---|---|---|
| E1 | **Sheri Nowmer** | **President** | Inaccuracy |
| E2 | Derrick Whelply | Store Manager | Accuracy |
| E3 | Michael Spence | VP Country Manager | Incompleteness |
| **E4** | **Kim Brunner** | **HQ Information Systems** | **Nonmember** |

Table 9. Actual Data Captured on Employee Table

The Cartesian product for Features and Employees (denoted by R) is shown in Table 10.
The incomplete set is denoted by $R_C$ and is shown in Table 11. Tuples in $R_C$ are of two types: (a) tuples that are products of a tuple from Feature$_C$ and a tuple from Employee$_C$, and (b) tuples that are products of an accurate or inaccurate tuple from Features (Employees) and a tuple from Employees$_C$ (Features$_C$).
Formally, let $C_1$ and $C_2$ be two cubes on which the Cubic product operation is performed, and let R be the result of the operation. Furthermore, let $t_1$ be a tuple in $C_1$ (or $C_{1C}$), $t_2$ be a tuple in $C_2$ (or $C_{2C}$), and t be a tuple in $R$ (or $R_C$). Table 12. summarizes how tuples should be categorized in $R$. Note that the concatenation of $t_1 \in C_{1N}$ and $t_2 \in C_{2C}$, and $t_1 \in C_{1C}$ and $t_2 \in C_{2N}$, are not meaningful to our analysis because they appear neither in the true world of R nor in the observed version of $R$.

| Product_ID | Customer_ID | Store_Address | Store_Cost | Employee_ID | Employee_Name | Status |
|---|---|---|---|---|---|---|
| P1 | 334-1626-003 | 5203 Catanzaro Way | 10,031 | E1 | **Sheri Nowmer** | I |
| P1 | 334-1626-003 | 5203 Catanzaro Way | 10,031 | E2 | Derrick Whelply | A |
| P1 | 334-1626-003 | 5203 Catanzaro Way | 10,031 | **E4** | **Kim Brunner** | N |
| P2 | 334-1626-006 | **4 Valley View** | 8,277 | E1 | **Sheri Nowmer** | I |
| P2 | 334-1626-006 | **4 Valley View** | 8,277 | E2 | Derrick Whelply | I |
| P2 | 334-1626-006 | **4 Valley View** | 8,277 | **E4** | **Kim Brunner** | N |
| P5 | 334-1626-008 | 321 herry Ct. | 11,412 | E1 | **Sheri Nowmer** | N |
| P5 | 334-1626-008 | 321 herry Ct. | 11,412 | E2 | Derrick Whelply | N |
| P5 | 334-1626-008 | 321 herry Ct. | 11,412 | **E4** | **Kim Brunner** | N |

Table 10. The Cartesian Product Cube R

| Product_ID | Customer_ID | Store_Address | Store_Cost | Employee_ID | Employee_Name |
|---|---|---|---|---|---|
| P1 | 334-1626-003 | 5203 Catanzaro Way | 10,031 | E3 | Michael Spence |
| P2 | 334-1626-006 | **4 Valley View** | 8,277 | E3 | Michael Spence |
| P3 | **334-1626-012** | **234 Coit Rd.** | **8,230** | E1 | **Sheri Nowmer** |
| P3 | **334-1626-012** | **234 Coit Rd.** | **8,230** | E2 | Derrick Whelply |
| P3 | **334-1626-012** | **234 Coit Rd.** | **8,230** | E3 | Michael Spence |
| P4 | 334-1626-005 | 1250 Coggins Drive | 8,856 | E1 | **Sheri Nowmer** |
| P4 | 334-1626-005 | 1250 Coggins Drive | 8,856 | E2 | Derrick Whelply |
| P4 | 334-1626-005 | 1250 Coggins Drive | 8,856 | E3 | Michael Spence |

Table 11. Feature Class Cube C

| $C_1 \times C_2$ | $t_2 \in C_{2A}$ | $t_2 \in C_{2I}$ | $t_2 \in C_{2N}$ | $t_2 \in C_{2C}$ |
|---|---|---|---|---|
| $t_1 \in C_{1A}$ | $t \in R_A$ | $t \in R_I$ | $t \in R_N$ | $t \in R_C$ |
| $t_1 \in C_{1I}$ | $t \in R_I$ | $t \in R_I$ | $t \in R_N$ | $t \in R_C$ |
| $t_1 \in C_{1N}$ | $t \in R_N$ | $t \in R_N$ | $t \in R_N$ | – |
| $t_1 \in C_{1C}$ | $t \in R_C$ | $t \in R_C$ | – | $t \in R_C$ |

Table 12. Tuple for the Cubic product Operation

The cardinality of the accurate, inaccurate, and nonmember tuples in R, and the incomplete tuples in $R_C$, are as shown below.
The cardinality of the accurate, inaccurate, and nonmember tuples in R, and the incomplete tuples in $R_C$, are as shown below.

$$|R_A| = |L_{1A}| \bullet |L_{2A}| \tag{13}$$

$$|R_I| = |L_{1A}| \bullet |L_{2I}| + |L_{1I}| \bullet |L_{2A}| + |L_{1I}| \bullet |L_{2I}| \tag{14}$$

$$|R_N| = |L_{1A}| \bullet |L_{2N}| + |L_{1I}| \bullet |L_{2N}| + |L_{1N}| \bullet |L_{2A}| + |L_{1N}| \bullet |L_{2I}| + |L_{1N}| \bullet |L_{2N}| \tag{15}$$

$$|R_C| = |L_{1A}| \bullet |L_{2C}| + |L_{1I}| \bullet |L_{2C}| + |L_{1C}| \bullet |L_{2A}| + |L_{1C}| \bullet |L_{2I}| + |L_{1C}| \bullet |L_{2C}| \tag{16}$$

Let $\mathrm{Pr}_A(i), \mathrm{Pr}_I(i), \mathrm{Pr}_N(i)$ and $\mathrm{Pr}_C(i)$ indicate the quality risks of $S_i$ i =1, 2. $\mathrm{Pr}_A(R), \mathrm{Pr}_I(R), \mathrm{Pr}_N(R)$ and $\mathrm{Pr}_C(R)$ indicate the quality risks of the Cubic product R. Using $|R| = |R_1| \bullet |R_2|$ and the definitions in Section Cube-Level Risks, we have

$$|\mathrm{Pr}_A(R)| = \frac{|L_{1A}|}{|L_1|} \bullet \frac{|L_{2A}|}{|L_2|} = \mathrm{Pr}_A(C_1) \bullet \mathrm{Pr}_A(C_2) \tag{17}$$

$$\begin{aligned} |\mathrm{Pr}_I(R)| &= \frac{|L_{1A}| \bullet |L_{2I}| + |L_{1I}| \bullet |L_{2A}| + |L_{1I}| \bullet |L_{2I}|}{|L_1| \bullet |L_2|} \\ &= \mathrm{Pr}_A(C_1) \bullet \mathrm{Pr}_I(C_2) + \mathrm{Pr}_A(C_1) \bullet \mathrm{Pr}_I(C_1) + \mathrm{Pr}_I(C_1) \bullet \mathrm{Pr}_I(C_2) \end{aligned} \tag{18}$$

$$|\mathrm{Pr}_N(R)| = \frac{|L_{1A}| \bullet |L_{2N}| + |L_{1I}| \bullet |L_{2N}| + |L_{1N}| \bullet |L_{2A}|}{|L_1| \bullet |L_2|} + \frac{|L_{1N}| \bullet |L_{2I}| + |L_{1N}| \bullet |L_{2N}|}{|L_1| \bullet |L_2|}$$

$$\begin{aligned} &= \mathrm{Pr}_N(C_1) \bullet (1 - \mathrm{Pr}_N(C_2)) + \mathrm{Pr}_N(C_2) \bullet (1 - \mathrm{Pr}_N(C_1)) + \mathrm{Pr}_N(C_1) \bullet \mathrm{Pr}_N(C_2) \\ &= \mathrm{Pr}_N(C_1) + \mathrm{Pr}_N(C_2) - \mathrm{Pr}_N(C_1) \bullet \mathrm{Pr}_N(C_2) \end{aligned} \tag{19}$$

From equality (17), we have

$$\frac{|R_C|}{|R|} = (1 - \mathrm{Pr}_N(C_1)) \bullet (1 - \mathrm{Pr}_N(C_2)) \bullet \frac{\mathrm{Pr}_C(C_1) + \mathrm{Pr}_C(C_2) - \mathrm{Pr}_C(C_1) \bullet \mathrm{Pr}_C(C_2)}{(1 - \mathrm{Pr}_C(C_1)) \bullet (1 - \mathrm{Pr}_C(C_2))}$$

Therefore, we have

$$\mathrm{Pr}_C(R) = \frac{|R_C|/|R|}{1 - \mathrm{Pr}_M(R) + |R_C|/|R|}$$

$$\begin{aligned} &= \left[ (1 - \mathrm{Pr}_N(C_1)) \bullet (1 - \mathrm{Pr}_N(C_2)) \bullet \frac{\mathrm{Pr}_C(C_1) + \mathrm{Pr}_C(C_2) - \mathrm{Pr}_C(C_1) \bullet \mathrm{Pr}_C(C_2)}{(1 - \mathrm{Pr}_C(C_1)) \bullet (1 - \mathrm{Pr}_C(C_2))} \right] \bullet \\ &\left\{ 1 - (\mathrm{Pr}_N(C_1) + \mathrm{Pr}_N(C_2) - \mathrm{Pr}_N(C_1) \bullet \mathrm{Pr}_N(C_2)) \right. \\ &\left. + \left[ (1 - \mathrm{Pr}_N(C_1)) \bullet (1 - \mathrm{Pr}_N(C_2)) \bullet \frac{\mathrm{Pr}_C(C_1) + \mathrm{Pr}_C(C_2) - \mathrm{Pr}_C(C_1) \bullet \mathrm{Pr}_C(C_2)}{(1 - \mathrm{Pr}_C(C_1)) \bullet (1 - \mathrm{Pr}_C(C_2))} \right] \right\}^{-1} \\ &= \mathrm{Pr}_C(C_1) + \mathrm{Pr}_C(C_2) - \mathrm{Pr}_C(C_1) \bullet \mathrm{Pr}_C(C_2) \end{aligned} \tag{20}$$

From Equality (17), we can see that the accuracy of the output of the Cubic product operator is less than the accuracy of either of the input cubes, and that the accuracy can become very low if the participating tables are not of high quality. Nonmembership and incompleteness also increase for the output.

## 5. Reducing the information quality risk for a finance company

### 5.1 Introduction

This case was part of a project undertaken for an auto financing company (AFC) to predict the propensities of its customers to buy its profitable offerings. According to the framework proposed by Su et al (Su, et al., 2008; Su et al., 2009c), the work presented in this chapter would be classified as a 'Pragmatics' information quality risk assessment.

The quality risk was restricted to assessments of the data along the following three criteria.

- Accuracy risk: The extracted data had to be verified against the respective origins in the warehouse. The data in the warehouse were not assessed for accuracy.
- Completeness risk: It is a critical data quality attribute, in particular for data warehousing applications that draw upon multiple internal and external data sources.
- Consistency risk: The extracted data had to be consistent with the minimal information requirements for the project -as stipulated by the Project Regulation and as listed in the Information Requirements Document.

Aberrations in the data discovered in the course of the assessment were documented and submitted to the warehouse administrators. However, an evaluation of the warehouse data was beyond the immediate scope(D. J. Kim, et al., 2008).

The main contribution of this case is the development of quantitative models to confirm the information quality risks in decision support for this finance company.

### 5.2 Key components of risk

We will use the framework in knowledge intensive business services(Su & Jin, 2007) to briefly review the key components of company risk.

1. Internal environment is the organization's philosophy for managing risk (risk appetite and tolerance, values, etc.);
2. Objective setting identifies specific goals that may be influenced by risk events;
3. Event identification recognizes internal or external events that affect the goals;
4. Risk assessment considers the probability of an event and its impact on organizational goals;
5. Risk response determines the organization's responses to risk events such as avoiding, accepting, reducing, or sharing;
6. Control activities focus on operational aspects to ensure effective execution of the risk response
7. Information and communication informs stakeholders of relevant information;
8. Monitoring continuously evaluates the risk management processes;

For compliance-driven risk programs, information requirements play a central role in dictating the risk architecture. We provide a set of guidelines to this financial institution to perform risk-based capital calculations. To comply with these guidelines, AFC must show they have the data (and up to seven years of history) required to calculate risk metrics such as probability of accuracy, loss completeness and consistency, etc.

### 5.3 The quality risks

Upon examination of the Information requirements, and the associated extraction process, the focal points for the extraction process were identified as the following.

- Mappings: The data extraction required linking data from the business definitions, as identified by the Project Regulation, to their encoding for the warehouse. Quality risk required an examination of these mappings.
- Parallel extraction: The extraction process for certain data was identical across the twelve product categories. Information quality could be assessed through examination of such data for a single product category for a single month.
- Peculiar extraction: Certain data were peculiar to specific product categories. These data had to be examined individually for assurance of quality.

Risk assessment comprised comparison of the extracted data with the parent data in the warehouse, and risks on the code used in the extraction.

The risks on the 'mappings' were performed on the items in Table 13.

| | | |
|---|---|---|
| QR1 | Product identifiers | It was checked that the roll-ups from the granular product levels to the product categories were accurate. |
| QR2 | Transaction identifiers | It was checked that the transactions used to measure the relationships among the finance company and its customers were restricted to customer-initiated transactions. |
| QR3 | Time identifiers | It was checked that the usage of the time identifiers to collate data from the fact tables was consistent with the encoding. |
| QR4 | Monthly balances per product category | It was checked that the monthly balance for a certain customer in a certain product category was the sum of the balances for all the customer's accounts in that product category for the same month. |
| QR5 | Valid accounts per product category | It was checked that the number of accounts held by a given customer in a given product category for a given month was calculated correctly. |

Table 13. Mappings' assessed for quality

| | |
|---|---|
| QR6 | Loan limits. |
| QR7 | Days to maturity. |
| QR8 | Overdraft limits. |
| QR9 | Promotional pricing information. |
| QR10 | Life/Disability insurance indicators. |

Table 14. Peculiarly extracted' data assessed for quality

Once it was verified that the mappings, as identified by Table 13, and had been accurately interpreted, the quality risks on the 'common extraction' items corresponded to verifying their extraction for a single month in any given product category. The quality risks for the 'parallel extraction' items were performed on the following.

The quality risks for the 'peculiar extraction' items were shown in Table 14. The quality risks comprised the verification of the respective data as the cumulative over all the accounts held by the customer in the particular product category.

### 5.4 Quality risk assessment

The data mining analysis did not directly use all the variables listed in the information requirements. However, it is easily seen that the existence of inaccurate, null, inconsistency, and incomplete attribute values have a direct impact on the aggregate values. For instance, consider the following query on the Loans table shown in Table 15.

| Cust_ID | Prod_ID | Loans Date | Quantity | Loan Amount | Status |
|---|---|---|---|---|---|
| C1 | P1 | 10-Mar-06 | 1000 | 100,031 | A |
| C1 | P1 | 22-apr-05 | 2000 | 76,342 | A |
| C2 | P2 | 06-may-06 | **3000** | 95,254 | I |
| C3 | P1 | 12-jun-07 | | | C |
| C3 | P2 | 10-sep-08 | **1200** | **83,277** | I |
| C4 | P1 | 14-aug-08 | 3600 | 90,975 | A |
| C5 | **P2** | 15-apr-07 | 6400 | 82,230 | M |
| C6 | **P1** | 18-jul-07 | **2100** | **19,450** | I |
| C6 | P3 | 23-nov-08 | 7800 | **38,645** | I |

Table 15. Customer Loans Table

SELECT SUM (Loans Amt) FROM Loans WHERE Prod ID = 'P1'

The query returns 286798 for the aggregate sum value. This, however, is not the true value because a) the inaccurate value 19,450 deviates from the actual value of 19,206; b) the inconsistency value 6400 contributes to this aggregate while it should not; c) the existential null value does not contribute to the sum while its true value of 3500 should; d) the values of 5200 and 7800 in the incomplete data set do not contribute to the sum while they should. Accounting for all the errors, the true aggregate sum value for this query is 65,500 which deviates about 23% from the query result.

It is, therefore, essential that the number of inaccurate, existential null, inconsistency, and incomplete values for each attribute be obtained in order to adjust the query result for the errors caused by these values. Auditing every single value in a database or data warehouse table that typically contains very large numbers of rows and attributes is expensive and impractical. Instead, sampling strategies can be used to estimate these errors as described next.

#### 5.4.1 Strategies for reducing risk

In order to estimate the number of inconsistency, we draw a random sample without replacement from the set of identifier attributes of L and verify the number of accurate and inaccurate values; denoted by $n_{k:A}$ and $n_{k:I}$, respectively; in the sample as shown in Fig. 3.

Let |L| denote the cardinality of L; let $n_k$ be the sample size; and let $l_{k:A}$ be the total number of accurate identifiers in L that must be estimated. The maximum likelihood estimator (MLE) of $l_{k:C}$, denoted by $\hat{l}_{k:C}$, is an integer that maximizes the probability distribution of the accurate identifiers in *L*. This probability follows a hypergeometric distribution given by:

| | | |
|---|---|---|
| $n_k$ | $K(k_1,...,k_m)$ | |
| $n_{k:C}$ | $\forall v_{ki} \leftarrow C; i \in \{1,...,m\}$ | |
| $n_{k:I}$ | $\exists\ \forall v_{ki} \leftarrow I; i \in \{1,...,m\}$ | Inconsistency |

Fig. 3. Identifier sampling

$$p(n_{k:A} = x) = \frac{\binom{L_{k:A}}{x}\binom{|L| - l_{k:A}}{n_k - x}}{\binom{|L|}{n_k}} \tag{21}$$

Using the closed form expression we have:

$$\hat{l}_{k:A} = \left\lceil \frac{n_{k:A}(|L|+1)}{n_k} \right\rceil \tag{22}$$

where $\lceil \cdot \rceil$ is the ceiling for any given number. The MLE for the inaccurate identifiers in $L$ (i.e., inconsistencys), denoted by $\hat{l}_{k:M}$ is then given by:

$$\hat{l}_{k:M} = |L| - \hat{l}_{k:A} = |L| - \left\lceil \frac{n_{k:A}(|L|+1)}{n_k} \right\rceil \tag{23}$$

In non-identifier attribute sampling, as shown in Fig. 4.

| | | |
|---|---|---|
| $K(k_1,...,k_m)$ | $q_i\ i \in \{1,...,n\}$ | $n_q$ |
| $\forall v_{ki} \leftarrow A; i \in \{1,...,m\}$ | $v_{qi} \leftarrow A$ | $n_{q:A}$ |
| | $v_{qi} \leftarrow I$ | $n_{q:I}$ |
| | $v_{qi} \leftarrow N$ | $n_{q:N}$ |
| $\forall v_{ki} \leftarrow I; i \in \{1,...,m\}$ | $v_{qi} \leftarrow A$ | Incompleteness |
| | $v_{qi} \leftarrow I$ | |
| | $v_{qi} \leftarrow N$ | |

Fig. 4. Non-identifier sampling.

The corresponding identifier values are also retrieved since the non-identifier attribute values find their meaning only in conjunction with their corresponding identifiers.

Let $l_{q:A}$, $l_{q:I}$, and $l_{q:N}$ be the total numbers of accurate, inaccurate, and existential null values in $q_i$ with an accurate identifier that need to be estimated. Their MLEs, denoted by $\hat{l}_{q:A}$, $\hat{l}_{q:I}$, and $\hat{l}_{q:N}$, are integers that maximize the probability distribution of these attribute value types in $q_i$. This probability function follows a multivariate hyper geometric distribution given by

$$p(n_{q:A}=x, n_{q:I}=y, n_{q:N}=z)=\frac{\binom{L_{q:A}}{x}\binom{L_{q:I}}{y}\binom{L_{q:N}}{z}}{\binom{\hat{l}_{k:A}}{n_q}} \tag{24}$$

A good approximation of MLEs can be obtained by assuming that $l_{q:A}$, $l_{q:I}$, and $l_{q:N}$ are integral multiples of $n_q$. Their estimates are then given by

$$\hat{l}_{q:A}=\left\lceil \frac{n_{q:A}(\hat{l}_{k:A}+1)}{n_q} \right\rceil ; \hat{l}_{q:I}=\left\lceil \frac{n_{q:I}(\hat{l}_{k:A}+1)}{n_q} \right\rceil ; \hat{l}_{q:N}=\left\lceil \frac{n_{q:N}(\hat{l}_{k:A}+1)}{n_q} \right\rceil \tag{25}$$

We propose using the Simple-Recapture sampling method to obtain an assessment for the size of the incomplete data set $L_C$. For this purpose, we assume that $|L|$ tuples have been sampled from T and $\hat{l}_{k:A}$ is obtained and this sampling has been done twice. The MLE estimates for $|L|$ and $|L_C|$ are then given by:

$$\left|\hat{T}\right|=\frac{|L|^2}{\hat{l}_{k:A}};\ |L_C|=\left|\hat{T}\right|-|L|-\hat{l}_{k:M}=\frac{|L|^2}{\hat{l}_{k:A}}-\hat{l}_{k:A} \tag{26}$$

### 5.4.2 COUNT

COUNT is used to retrieve the cardinality of L or it functions on one of the identifier attributes, the true COUNT, denoted by $COUNT^T$, is the number of tuples with accurate identifiers plus the cardinality of the incomplete set:

$$COUNT^T(k_i)=\hat{l}_{k:A}+\left|\hat{L}_C\right| \tag{27}$$

When COUNT operates on one of the non-identifier attributes, the true count is the sum of accurate, inaccurate, and incomplete values:

$$COUNT^T(q_i)=\hat{l}_{q:A}+\hat{l}_{q:I}+\hat{l}_{q:N}+\left|\hat{L}_C\right| \tag{28}$$

### 5.4.3 SUM

The distributions of attribute value types within their underlying domains affect the assessment of the true SUM value. Therefore, we assume that the attribute value types could have a uniform distribution depending on the error generating processes. We use $\bar{\lambda}_{k:A}$ for each value in the incomplete data set and the estimated true sum will be given by

$$SUM^T(k_i)=\bar{\lambda}_{k:A}(\hat{l}_{k:A}+\left|\hat{L}_C\right|) \tag{29}$$

When SUM operates on a non-identifier attribute, the estimate for the true SUM value can be obtained by substituting the inaccurate, existential nulls and incomplete values with $\bar{\lambda}_{q:A}$ which is given by

$$SUM^T(q_i)=\bar{\lambda}_{k:A}(\hat{l}_{q:A}+\hat{l}_{q:I}+\hat{l}_{q:N}+\left|\hat{L}_C\right|) \tag{30}$$

### 5.4.4 AVERAGE

The estimated true value returned by the AVERAGE function on an identifier (non-identifier) attribute is given by the ratio of the estimated true SUM and true COUNT:

$$\text{AVERAGE}^{T}(\mathrm{k_i}) = \frac{SUM^{T}(k_i)}{COUNT^{T}(k_i)} = \bar{\lambda}_{k:A} \tag{31}$$

$$\text{AVERAGE}^{T}(\mathrm{q_i}) = \frac{SUM^{T}(q_i)}{COUNT^{T}(q_i)} = \bar{\lambda}_{q:A} \tag{32}$$

### 5.5 Quality risk initiatives

We present the nine key steps to successful deployment of an information quality program for a risk management initiative.

1. Identify the information elements necessary to manage credit risk. Identifying all the information elements and sources necessary to calculate company risk is no mean feat. Risk data such as QR1, QR2… QR10, for example, can each require the identification of several different product identifiers.
2. Define a information quality measurement framework.

The key dimensions that data quality traditionally measures include consistency (21), completeness (24), conformity, accuracy(26), duplication(28), and integrity(30). In addition, for risk calculations, dimensions such as continuity, timeliness, redundancy, and uniqueness can be important.

3. Institute an audit to measure the current quality of information.

Perform an information quality audit to identify, categorize, and quantify the quality of information based upon the decisions made in the previous step.

4. Define a target set of information quality metrics against each attribute, system, application, and company.

Based on the audit results and the impact that each attribute, application, database, or system will have on the ability of your organization to manage risk, the organization should define a set of information quality targets for each attribute, system, application, or company.

5. Set up a company wide information quality monitoring program, and use data to drive process change.
6. Identify gaps against targets.

The quality risks on the data discovered the following critical gaps.

| QR1 | **QR2** | QR3 | QR4 | QR5 | QR6 | **QR7** | QR8 | **QR9** |
|---|---|---|---|---|---|---|---|---|
| 0.6 | **0.4** | 0.8 | 0.5 | 0.8 | 0.6 | **0.4** | 0.8 | **0.2** |

Table 16. Critical Gaps

The issues listed in Table 16 after verification with the finance company analysts and the warehouse administrators.

Other quality issues were unpopulated data fields and unary data. In each case, these gaps were communicated to the warehouse, but were considered non-critical and did not require immediate address.

### 5.6 Remarks

The main contribution of this work is the illustration of a quantitative method that condensed the task of verifying the credit data to ten quality checks. The quality checks listed here can be transferred to other prediction analyses with a few modifications. However, their categorization as 'mappings', 'parallel extraction' and 'peculiar extraction' is a general, transferable framework. This proposition is elucidated in a methodology below.

We have provided a formal definitions of attribute value types (i.e., accurate, inaccurate, consistency, and incomplete) within the data cube model. Then, we presented sampling strategies to determine the maximum likelihood estimates of these value types in the entire data population residing in data warehouses. The maximum likelihoods estimates were used in our metrics to estimate the true values of scalars returned by the aggregate functions. This study can further be extended to estimate the IQ by the widely used Group By clause, partial sum, and the OLAP functions.

## 6. Case study for medical risk management

This section describes blood stream infection; we analyzed the effects of lactobacillus therapy and the background risk factors of bacteria detection on blood cultures. For the purpose of our study, we used the clinical data collected from the patients, such as laboratory results, isolated bacterium, anti-biotic agents, lactobacillus therapy, various catheters, departments, and underlying diseases.

### 6.1 Mathematical model

We propose entropy of clinical data to quantify the information quality. The entropy of clinical data is derived through modeling the clinical data as Joint Gaussian Random Variables (JGRVs) and applying the exponential correlation models that are verified by experimental data. We prove that a simple yet Effective Asynchronous Sampling Strategy (EASS) is able to improve the information quality of clinical data by evenly shifting the sampling moments of nodes from each other. At the end of this section, we derive the lower bound on the performance of EASS to evaluate its effectiveness on improving the information quality.

#### 6.1.1 Entropy of clinical data

Without loss of generality, we assume clinical data from n different locations in the monitored area are JGRVs with covariance matrix C, whose element, $c_{i,j}$ , is given in the following:

$$c_{ij} = \begin{cases} \sigma_i^2 & i = j, for\ i \le n, j \le n, \\ \sigma_i \sigma_j \Pr_{i,j} & i \ne j, for\ i \le n, j \le n, \end{cases}$$

where $\sigma_i$ and $\sigma_j$ are the standard deviation of the clinical data $S_i$ and $S_j$ , respectively. Normalizing the covariance matrix leads to the correlation matrix A, which consists of the correlation coefficients of clinical data. The entry of A, $a_{i,j}$ , is given as follows:

$$a_{ij} = \begin{cases} 1 & i = j, for\ i \le n, j \le n, \\ \Pr_{i,j} & i \ne j, for\ i \le n, j \le n, \end{cases} \quad (33)$$

Then, according to the definition of entropy of JGRVs, the entropy of the clinical data, H, is

$$H = \frac{1}{2}\log(2\pi e)^n \det C - \log\Delta \tag{34}$$

Where $\log\Delta$ is a constant due to quantization. detC is the determinant of the covariance matrix, which is:

$$\det C = \prod_{i=1}^{n} \sigma_i^2 \det A \tag{35}$$

For the sake of simplicity, we do not elaborate on the closed-form expression of the entropy. However, we will show, in the following, how to improve the information quality through increasing the entropy of clinical data.

### 6.1.2 Quality improvement

In the discussion on correlation model, we show that asynchronous sampling is able to produce less correlated data compared with synchronous sampling. With the entropy model based on correlation coefficients, the following discussion further explains that the information quality of clinical data improves through asynchronous sampling. Here, we quantify the information quality using entropy of the clinical data. Then, we need to prove $H \le \hat{H}$ , where $\hat{H}$ is the entropy with respect to asynchronous sampling and H is that of synchronous sampling. Therefore, we have the following theorem and its proof:

$$H \le \hat{H} \tag{36}$$

$$H = \frac{1}{2}\log(2\pi e)^n \prod_{i=1}^{n} \sigma_i^2 \det A - \log\Delta \tag{37}$$

$$\hat{H} = \frac{1}{2}\log(2\pi e)^n \prod_{i=1}^{n} \sigma_i^2 \det \hat{A} - \log\Delta \tag{38}$$

As the entropy of sensory data increases after applying asynchronous sampling, we conclude that asynchronous sampling is able to improve the information quality of sensory data if the sensory data are temporal-spatial correlated.

### 6.1.3 Asynchronous sampling strategy

Through quantifying the information quality of sensory data, we show that asynchronous sampling can improve information quality by introducing non-zero sampling shifts. Instead of maximizing entropy through asynchronous sampling, we propose EASS that assigns equal sampling shifts to different locations. Given a set of sensors taking samples periodically, the sampling moments of the ith sensor is ti, ti + T, ti + 2T, … , where T is the sampling interval of the sensor nodes. Accordingly, we define the time shifts for sensor nodes, _i, as follows:

$$T_i = \begin{cases} t_{i+1} - t_i & i = 1,...,n-1, \\ T + t_1 - t_i & i = n \end{cases}$$

Thus we have

$$\sum_{k=1}^{n} \tau_k = T \tag{39}$$

For the proposed EASS, $\tau_i = \frac{T}{n}$ , for $i \le n$ .

Table 17 shows all the components of this dataset. The following decision tree shown in Fig. 5.was obtained as the relationship between the bacteria detection and the various factors, such as diarrhea, lactobacillus therapy, antibiotics, surgery, tracheotomy, CVP/IVH catheter, urethral catheter, drainage, other catheter. Fig. 5. shows the sub-tree of the decision tree on lactobacillus therapy = Y (Y means its presence.)

| Item | Attributes |
|---|---|
| Patient's Profile | ID, Gender, Age |
| Department | Department, Ward, Diagnosis |
| Order | Background Diseases, Sampling Date, Sample, No. |
| Symptom | Fever, Cathether(5), Traheotomy, Endotracheal intubation, Drainage(5) |
| Examination Data | CRP, WBC, Urin data, Liver/Kidney Function, Immunology |
| Therapy | Antibiotic agents(3), Steroid, Anti-cancer drug, Radiation Therapy, Lactobacillus Therapy |
| Culture | Colony count, Bacteria, Vitek biocode, $\beta$−lactamase |
| Susceptibility | Cephems, Penicillins, Aminoglycoside, Macrolides, Carbapenums, Chloramphenicol,Rifanpic, VCM, etc. |

Table 17. Attributes in a Dataset on Infection Control

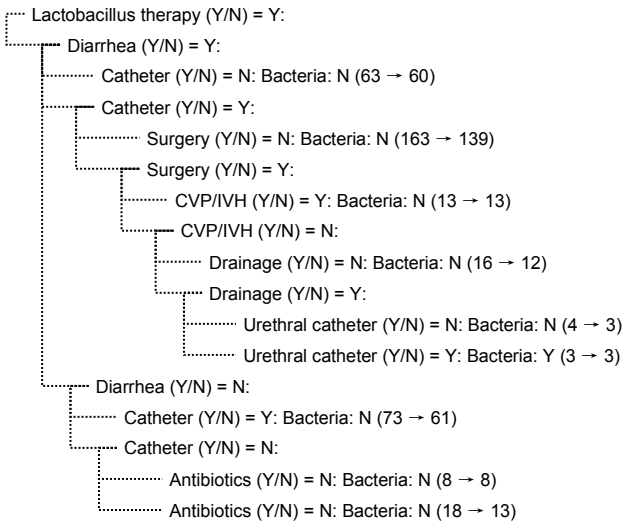


Fig. 5. Sub-tree on lactobacillus therapy(Y/N) = Y

### 6.2 Discussion and conclusion

Our methods can be used in hospital information system (HIS) analysis environments to determine how source data of different quality could impact medical databases derived using selection, projection, and Cartesian product operations. There was a lack of insight in which element of medical information quality (MIQ) was most relevant and a lack of insight into how implications of MIQ could be quantified. Our method would be useful in identifying which data sets will have acceptable quality, and which one will not. Based on this chapter four conclusions can be drawn:

- The formulation of the conceptual and mathematical model is general and therefore widely applicable.
- The model provides risk detection discovers patterns or information unexpected to domain experts
- The model can be used to a new cycle of risk mining process
- Three important process: risk detection, risk clarification and risk utilization are proposed.

The case study illustrated that the model could be parameterized with data collected from contractors through a database. Once parameterized with acceptable preciseness, applications valuable for society may be expected.

## 7. Conclusions

Our analysis can be used in business data mining environments to determine how source data of different quality could impact those DM derived using Restriction, Projection, and Cubic product operations. Because business data mining could support multiple such applications, our analysis would be useful in identifying which data sets will have acceptable quality, and which ones will not. Finally, our results can be implemented on top of data warehouses engine that can assist end users to obtain quality risks of the information they receive. The quality information will allow users to account for the reliability of the information received thereby leading to decisions with better outcomes.

## 8. Acknowledgment

We would like to thank NNSFC (National Natural Science Foundation of China) for supporting Ying Su with a project (70772021, 70831003).

## 9. References

Ballou, D.P., & Pazer, H.L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Management Science, 31(2), 150.

Bose, I., & Mahapatra, R.K. (2001). Business Data Mining - A Machine Learning Perspective. Information & Management, 39(3), 211-225.

Chen, S.Y., & Liu, X.H. (2004). The contribution of data mining to information science. Journal of Information Science, 30(6), 550-558.

Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., & Kechadi, T. (2007). Exploratory spatio-temporal data mining and visualization. Journal of Visual Languages and Computing, 18(3), 255-279.

Cowell, R.G., Verrall, R.J., & Yoon, Y.K. (2007). Modeling Operational Risk with Bayesian Networks. Journal of Risk and Insurance, 74(4), 795-827.

DeLone, W.H., & McLean, E.R. (2003). The DeLone and McLean model of information systems success: a ten-year update. Journal of Management Information Systems, 19(4), 9-30.

English, L.P. (1999). Improving data warehouse and business information quality methods for reducing costs and increasing profits New York: Wiley.

Eppler, M.J. (2006). Managing information quality increasing the value of information in knowledge-intensive products and processes (2nd ed.). New York: Springer.

Fisher, C.W., Chengalur-Smith, I., & Ballou, D.P. (2003). The impact of experience and time on the use of Data Quality Information in decision making. Information Systems Research, 14(2), 170-188.

Goodhue, D.L. (1995). Understanding user evaluations of information systems. Management Science, 41(12), 1827.

Hand, D.J., Mannila, H., & P. Smyth. (2001). Principles of Data Mining: MIT Press.

Huang, K.-T., Lee, Y.W., & Wang, R.Y. (1999). Quality information and knowledge. Upper Saddle River, N.J. : Prentice Hall PTR.

Jin, R., Vaidyanathan, K., Ge, Y., & Agrawal, G. (2005). Communication and Memory Optimal Parallel Data Cube Construction. IEEE Transactions on Parallel & Distributed Systems, 16(12), 1105-1119.

Kim, D.J., Ferrin, D.L., & Rao, H.R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems, 44(2), 544-564.

Kim, W., Choi, B.J., Hong, E.K., Kim, S.K., & Lee, D. (2003). A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery, 7(1), 81-99.

Michalski, G. (2008). Operational Risk in Current Assets Investment Decisions: Portfolio Management Approach in Accounts Receivable. Agricultural Economics-Zemedelska Ekonomika, 54(1), 12-19.

Mitra, P., & Chaudhuri, C. (2006). Efficient algorithm for the extraction of association rules in data mining. Computational Science and Its Applications - Iccsa 2006, Pt 2, 3981, 1-10.

Mohamadi, H., Habibi, J., Abadeh, M.S., & Saadi, H. (2008). Data mining with a simulated annealing based fuzzy classification system. Pattern Recognition, 41(5), 1824-1833.

Mucksch, H., Holthuis, J., & Reiser, M. (1996). The Data Warehouse Concept - An Overview. Wirtschaftsinformatik, 38(4), 421-&.

Sen, A., & Sinha, A.P. (2007). Toward Developing Data Warehousing Process Standards: An Ontology-Based Review of Existing Methodologies. IEEE Transactions on Systems, Man & Cybernetics: Part C - Applications & Reviews, 37(1), 17-31.

Sen, A., Sinha, A.P., & Ramamurthy, K. (2006). Data Warehousing Process Maturity: An Exploratory Study of Factors Influencing User Perceptions. IEEE Transactions on Engineering Management, 53(3), 440-455.

Shao, T., & Krishnamurty, S. (2008). A clustering-based surrogate model updating approach to simulation-based engineering design. Journal of Mechanical Design, 130(4), -.

Su, Y., & Jin, Z. (2006). A Methodology for Information Quality Assessment in the Designing and Manufacturing Process of Mechanical Products. In L. Al-Hakim (Ed.), Information Quality Management: Theory and Applications (pp. 190-220). USA: Idea Group Publishing.

Su, Y., & Jin, Z. (2007, September 21-23). In Assuring Information Quality in Knowledge intensive business services (Vol. 1, pp. 3243-3246). Paper presented at the 3rd International Conference on Wireless Communications, Networking, and Mobile Computing (WiCOM '07), Shanghai, China. IEEE Xplore.

Su, Y., Jin, Z., & Peng, J. (2008). Modeling Data Quality for Risk Assessment of GIS. Journal of Southeast University (English Edition), 24(Sup), 37-42.

Su, Y., Peng, G., & Jin, Z. (2009a, September 20 to 22). In Reducing the Information Quality Risk in Decision Support for a Finance Company. Paper presented at the International Conference on Management and Service Science (MASS'09), Beijing, China. IEEE Xplore.

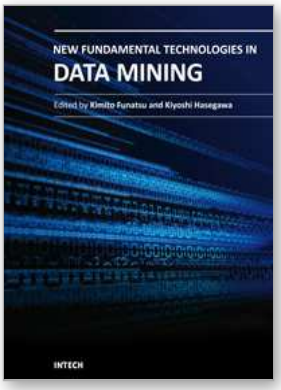
Su, Y., Peng, J., & Jin, Z. (2009b, December 18-20). In Modeling Information Quality for Data Mining to Medical Risk Management (pp. 2336-2340). Paper presented at the The 1st International Conference on Information Science and Engineering (ICISE2009), Nanjing,China. IEEE.

Su, Y., Peng, J., & Jin, Z. (2009c). Modeling Information Quality Risk for Data Mining in Data Warehouses. Journal of Human and Ecological Risk Assessment, 15(2), 332 - 350.

Wand, Y., & Wang, R.Y. (1996). Anchoring data quality dimensions in ontological foundations. Association for Computing Machinery. Communications of the ACM, 39(11), 86-95.

Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5.

Zmud, R.W. (1978). AN EMPIRICAL INVESTIGATION OF THE DIMENSIONALITY OF THE CONCEPT OF INFORMATION. Decision Sciences, 9(2), 187-195.




**New Fundamental Technologies in Data Mining**

Edited by Prof. Kimito Funatsu

ISBN 978-953-307-547-1
Hard cover, 584 pages
**Publisher** InTech
**Published online** 21, January, 2011
**Published in print edition** January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by "Data Mining" address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

**How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Ying Su (2011). Modeling Information Quality Risk for Data Mining and Case Studies, New Fundamental Technologies in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, InTech, Available from: http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/modeling-information-quality-risk-for-data-mining-and-case-studies

INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821