A GENERAL RANKING STRATEGY FOR DATA ACCURACY MANAGEMENT

(Research-in-Progress)

Irit Askira Gelman
DQIQ
Askira@cox.net

Abstract: A series of recent studies proposed a construct named *damage* and a set of models for estimating the damage in a chosen class of information systems. The perception that underlies the proposed construct is that, all other things being equal, it would be beneficial to assign priority to the elimination of data errors that have a stronger negative effect on output accuracy (i.e., output accuracy is lower) over data errors that have a weaker effect. In this paper we extend the work on damage by considering its use with information systems in general, rather than a specific class of information systems. Mainly, we propose a general strategy for ranking the inputs according to the damage that errors in each input inflict on the output of the system. A major advantage of this strategy is that it focuses the ranking effort on a subcomponent of the information system that can be substantially smaller and simpler than the information system as a whole.

Key Words: Accuracy, Ranking Strategy, Data Quality, Data Quality Management

INTRODUCTION

The overall annual cost of poor data quality to businesses in the US has been estimated in the hundreds of billions of dollars (Eckerson 2002) and the overall cost to individual organizations is believed to be 10%-20% of their revenues (Redman 2004). However, these estimates are not impressive enough, apparently, to drive organizations to action. For instance, most organizations have no plans for improving data quality in the future (Eckerson 2002). In the face of this neglect, there is a mounting conviction among both practitioners and researchers that an understanding of the economic aspect of data quality can be crucial for convincing organizations to increase their data quality efforts. An understanding of the economics of data quality can guide decisions on how much to invest in the quality of their information and how to allocate limited organizational resources (Wang and Strong 1996).

This paper is a product of a research project that addresses the need for models to support information quality resource allocation and design decisions. In particular, this research considers a key dimension of information quality, namely, accuracy (Wang and Strong 1996). Accuracy is defined as the degree to which the data or information are in conformance with the true values. Broadly, the questions that are of interest in this research project include, for instance:

(1) Assuming an information system that utilizes a specified set of input sources for producing required information, how can we identify the input sources that would yield the highest gain in information accuracy if their accuracy is improved? How can we identify the input sources that would offer the highest economic return if their accuracy is improved?

(2) How can we quantify the gain in information accuracy that would result from improving the accuracy of a chosen data source, and the subsequent economic return?

A series of studies (Askira Gelman 2009; Askira Gelman 2010; Askira Gelman Forthcoming) approached these questions by proposing a construct named *damage*, and a set of models for estimating the damage (Hevner et al. 2004; March and Smith 1995). The perception that underlies the proposed construct is that, all other things being equal, it would be beneficial to assign priority to the elimination of input errors that have a stronger negative effect on output accuracy

(i.e., output accuracy is lower) over input errors that have a weaker effect. In practical settings where, often, not all other things are equal, an estimate of the damage should be weighed by, or combined with, values of other relevant factors (e.g., the cost of higher accuracy), in order to yield a more comprehensive evaluation of an investment in the accuracy of a chosen input.

Work on the new concept of damage has focused, so far, on a single class of information systems. Systems in that class employ binary, multi-criteria decision or judgment rules that consist of conjunction and disjunction operations. Instances of such systems include databases, expert systems that utilize domain knowledge in the form of multi-criteria satisficing decision rules, and classification models that use decision trees for their purpose. (Askira Gelman 2009; Askira Gelman 2010), in particular, developed and evaluated a simple model for ranking the inputs of multi-criteria satisficing decision rules according to the damage that errors in each inflict. Askira Gelman (forthcoming) developed and evaluated a broad model for quantifying the damage in similar applications.

In this paper we extend the work on damage by considering its use with information systems in general rather than a specific class of information systems. Mainly, we propose a general strategy for ranking the inputs of an information system according to the damage that errors in each input inflict on the output of the information system. A major advantage of that strategy is that it focuses the ranking effort on a subcomponent of the information system that can be substantially smaller and simpler than the information system as a whole.

The next section provides a summary of relevant literature. Later, the section "Damage and Damage Ranking" defines the notion of damage and briefly discusses the potential value of a ranking of damage. The section "Theoretical Foundation of the Damage Ranking Strategy" introduces the theoretical basis that serves the proposed strategy. Finally, the section "Illustration: Application of the Ranking Strategy" demonstrates the application of the theory, and subsequent ranking strategy, in the case of information systems that employ satisficing decision rules as described above.

RELATED RESEARCH

An implicit assumption of this inquiry is that errors can be differentiated based on the intended use of the data. Counter to an approach that does not differentiate between errors (e.g., Janson, 1988; Parsaye and Chignel, 1993), an approach that differentiates between errors based on the intended use of the data is consistent with the currently accepted definition of data quality as "fitness for use." The concept of fitness for use emphasizes the context of the data, mainly the uses, users, and suppliers of the data (Juran, 1988).

A recent work that somewhat resembles the viewpoint that underlies our research introduces a data quality assessment method for database settings that accounts for variations in the potential utility of the data (Even and Shankaranarayan, 2007). In general, nowadays there are various tools and methods that guide the design from a data utilization perspective. For instance, Ballou and Pazer (Ballou and Pazer, 1985) belong in this class. They propose a framework for tracking numeric data errors through an information system to assist with estimating the impact of errors on the output. Notably, (Ballou et al. 1998) and several other studies (e.g., Shankaranarayan, Zaid, and Wang, 2003) have extended the model of Ballou and Pazer in several directions.

A few frameworks address the relationship between the quality of the raw data and the quality of the output of a relational database query. Parssian et al. (2004) assess the relationship between the quality of the data and the quality of the output of a query. The quality dimensions of interest in that study are completeness, accuracy, and membership. Motro and Rakov (1997) describe a data analysis method that identifies data subsets which are homogeneous in their soundness or

completeness. They employ aggregates of the data quality estimates that their method generates to assess the quality of query answers. Additional instances of work that accounts for the relationship between the quality of raw data and the quality of the output of queries include (Wang, Reddy, and Kon, 1995; Naumann et al., 1999; Avenali et al., 2008).

The contribution of our work on the concept of damage beyond previous work on the relationship between input accuracy and output accuracy lies in the special emphasis of this concept. The notion of damage is designed to assist in prioritization and resource allocation tasks in data quality management. As explained earlier, this paper, in particular, extends the work on damage to information systems at large.

Prioritization of data quality issues according to users' perceptions and needs is assisted by several methods and tools (e.g., Wang and Strong, 1996; Lee, Strong, Kahn, and Wang, 2002). Some tools are available today that assist directly with prioritization and resource allocation in data quality management settings (e.g. Ballou and Pazer, 1989). However, the relationship between input accuracy and output accuracy is neglected by these tools.

DAMAGE AND DAMAGE RANKING

The notion of damage is a central concept of this research. The damage that errors in an input inflict on output accuracy is defined as the change in output accuracy due to a change in the accuracy of that input. The idea that motivates this construct is that, all other things being equal, it would be beneficial to assign priority to the improvement of the accuracy of an input where accuracy deficiencies have a higher negative effect on output accuracy over an input where accuracy deficiencies have a less negative effect. For instance, assume that accuracy is measured by error rate. Suppose that, by decreasing the error rate in one of the information system's inputs by 1%, we decrease the error rate of the system's primary output by 0.5%, while a decrease in the error rate of a second input by 1% decreases the output error rate by 0.05%. Obviously, all other things being equal, it would be more effective to decrease the error rate of the first input than the second.

Technically, we view an information system as a function (e.g., Hamilton and Zeldin 1978; Wand and Weber, 1990). Likewise, we view the accuracy of the output of an information system as a function, such that input accuracy is one of its arguments. In general, input accuracy is not the only determinant of output accuracy. The nature of the process, processing errors, the correct input values, dependencies between processing errors and the correct input values, dependencies between input errors, dependencies between input errors and the correct input values, and dependencies between correct values, are important as well (Ballou and Pazer 1985; Askira Gelman 2004; Fisher et al. 2009). In the illustrative example of this paper, the accuracy of the output is derived analytically as a function of several such arguments. However, in the general case, determination of such a function is difficult at best. Practically, as we can see from the discussion below, the details of this function are not an absolute pre-requisite of our approach.

Definition (Damage): Let a^v and a^o denote the accuracy of input *v* and output *o*, respectively, of an information system *s*. The damage of errors in *v* to the accuracy of *o*, denoted by $\frac{\partial a^o}{\partial a^v}$, is the change in the accuracy of *o* due to a change in the accuracy of *v* when all other arguments of a^o are held fixed.

Our definition of damage avoids the use of a specific measure of accuracy. This choice is consistent with the intended generality of the proposed approach, i.e., the proposed strategy is independent of the preferred measure of accuracy.

An estimate of damage can be produced empirically, analytically, or through another computational method. An analytical method would derive the damage as a function of a set of pre-specified parameters. An empirical strategy can produce estimates of damage values through simulation of the information system. Alternatively, we can manipulate the accuracy of the inputs of an actual system, e.g., by adding input errors and monitoring the change in the accuracy of the system's output. A potential advantage of an approach that utilizes an existing system is that it can avoid the need to study many of the specifics of the accuracy function. However, an investigation of error distributions or dependency patterns is recommended before adding input errors, since such dependencies can affect the outcome. Clearly, estimates as described here can be prohibitively costly or practically impossible, especially when the information system is large or complex.

Rather than focus on a full-fledged damage estimate, this paper considers a ranking of the damage, i.e., an ordering of the inputs by the damage that the respective input errors inflict. A disadvantage of a ranking compared to a full-fledged quantitative measure is that a ranking may not enable a comprehensive assessment, i.e., an assessment that accounts for the damage as well as other relevant factors. We argue, on the other hand, that the damage ranking of two inputs can be useful per se, and, in addition, it may be obtained significantly more easily than quantitative damage values. Notably, the emphasis of this work is on relative ease and simplicity.

THEORETICAL FOUNDATION OF THE DAMAGE RANKING STRATEGY

The process of ranking a set of inputs of an information system with respect to the damage that the errors in each input inflict on an output can be viewed as a sequence of repeated ranking activities, each involving a single pair of inputs. Therefore, when discussing our ranking strategy, we can limit the discussion to a single input pair. We will denote the respective inputs v_i and v_j. Subsequently, the problem of interest here is how to easily rank inputs v_i and v_j of an information system *s* according to the damage that each generates to an output *o* of *s*.

Notation	Explanation
s	an information system; technically, *s* is a function
s'	an information system that is a component of *s*; technically, we employ function composition
$v,\ v_i,\ v_j$	inputs of *s*
$o,\ o'$	an output of s, s', respectively
$a^v, a_i^v, a_j^v, a^o, a^{o'}$	accuracy; the superscripts and subscripts describe the relevant variable, e.g., a_i^v denotes the accuracy of v_i

Table 1: Symbols in use in the presentation of the theoretical foundation

In essence, the proposed damage ranking strategy is based on the understanding that, given v_i and v_j as above, their ranking is often independent of a large share of s, i.e., it is determined by s', a sub-component of s, which can be much smaller and simpler than s. Hence, our ranking strategy aims to identify such a minimal sub-component s' and focus the ranking efforts on that component. Assumably, the process of ranking the damage to o', the output of s', would be simpler and less costly than the process of ranking the damage to o, the output of s. The remaining part of this section establishes the theory that underlies the proposed strategy, while a following section will illustrate the application of the theory and subsequent ranking strategy.

The understanding that the ranking of two inputs is often independent of much of the system is derived from the observation that information systems are typically modular. Modular design has been advocated throughout the history of software, and is supported by numerous ideas, methods, tools, and products. Common software development terminology such as subroutines, modular programming, object oriented programming, component-based software development, service oriented architecture, and numerous other popular concepts, reflect various interpretations of the principle of modularity that have evolved over the years. Accordingly, Assumption 1 states that v_i and v_j are inputs of s', a subsystem of s, such that the damage of errors in the recorded value of v_i $(/v_j)$ to the accuracy of output o of s is equal to the product of the damage of errors in the recorded values of v_i $(/v_j)$ to the accuracy of output o' of s' and the damage of errors in the computed values of o' to the accuracy of o.

Assumption 1 (Decomposability): Let s denote an information system. Let v_i and v_j denote two inputs of s. Then, there exists an information system s' such that s' is a component of s, v_i and v_j are inputs of s' and o' is an output of s', and the following conditions hold true. Let a_i^v, a_j^v, $a^{o'}$, a^o, denote the accuracy of v_i, v_j, o', and o, respectively. Let $\frac{\partial a^o}{\partial a_j^v}, \frac{\partial a^o}{\partial a_i^v}, \frac{\partial a^o}{\partial a^{o'}}, \frac{\partial a^{o'}}{\partial a_j^v}, \frac{\partial a^{o'}}{\partial a_i^v}$, denote the damage of errors in v_j to the accuracy of output o, the damage of errors in v_i to the accuracy of o, the damage of errors in o' to the accuracy of o, the damage of errors in v_j to the accuracy of o', and the damage of errors in v_i to the accuracy of o', respectively. Then:

$$\frac{\partial a^o}{\partial a_i^v} = \frac{\partial a^o}{\partial a^{o'}} \cdot \frac{\partial a^{o'}}{\partial a_i^v} \tag{1}$$

$$\frac{\partial a^o}{\partial a_j^v} = \frac{\partial a^o}{\partial a^{o'}} \cdot \frac{\partial a^{o'}}{\partial a_j^v} \tag{2}$$

Now, suppose that Assumption 1 holds true. Suppose also that the "smaller" problem of ranking the damage that errors in each of v_i and v_j inflict on output o' of s' can be solved at a reasonable cost. (Again, such a solution may be empirical, analytical, etc.) In other words, we assume next that the problem of ranking inputs v_i and v_j according to the size of $\frac{\partial a^{o'}}{\partial a_i^v}$ and

$\frac{\partial a^{o'}}{\partial a_j^v}$ has a feasible solution. Since such a solution would refer to s', not to s, it would not automatically solve the ranking problem in the bigger system. Mainly, there would still be a need to clarify the nature of the link between a ranking of the damage to o' and the ranking of the damage to o. However, equations (1)-(2) imply that the link between a ranking of the damage to o' and the ranking of the damage to o is captured by $\frac{\partial a^{o}}{\partial a^{o'}}$, which is the damage of errors in o' to the accuracy of o. Hence, we need to understand this damage, mainly its sign. If the sign of $\frac{\partial a^{o}}{\partial a^{o'}}$ is positive, then, according to (1)-(2), the same ranking that holds true in s' would also be valid in s. If that sign is negative then the former ranking would be reversed. We argue, however, that the sign of $\frac{\partial a^{o}}{\partial a^{o'}}$ is captured by the widespread belief in "Garbage In Garbage Out" (GIGO): a higher input accuracy produces a higher output accuracy. To the extent that this belief is valid, it informs us that the sign of the damage of errors in o' to the accuracy of o is positive. Assumption 2 designates the belief in GIGO.

Assumption 2 (GIGO): $\frac{\partial a^{o}}{\partial a^{o'}} \geq 0$.

Apart from a universal belief in GIGO among non-scientists, scientists have typically embraced the popular belief in GIGO and have treated GIGO as an axiom. Originally coined in the computer industry, this acronym, which indicates a strong positive link between input accuracy and output accuracy, is nowadays commonly accepted. Recently, however, there is a growing literature that suggests various conditions in which GIGO does not hold true. Notably, the illustrative example that we present in the next section has been associated with deviations from GIGO (Askira Gelman 2004, Askira Gelman 2007). Clearly, these deviations, as well as various related discoveries, prove that GIGO should not be taken to be true at all times. However, with this caveat in mind, we choose to utilize GIGO due to our conviction that GIGO is often enough valid, such that a strategy based on GIGO should not be discarded because of the deviations.

We can conclude that, if Assumptions 1 and 2 hold true, a ranking of the damage to o is *the same as* the respective ranking of the damage to o'. Proposition 1 states that, when Assumption 1 and Assumption 2 hold true, if the damage of errors in v_j to o' is higher than the damage of errors in v_i to o' then the damage of errors in v_j to o is higher than the damage of errors in v_i to o.

***Proposition* (Damage Ranking)**: Let v_i and v_j denote two inputs of s, consistent with Assumption 1 and Assumption 2. Then, $\frac{\partial a^{o'}}{\partial a_j^v} \geq \frac{\partial a^{o'}}{\partial a_i^v} \Rightarrow \frac{\partial a^{o}}{\partial a_j^v} \geq \frac{\partial a^{o}}{\partial a_i^v}$.

This proposition follows directly from Assumption 1 and Assumption 2.

The conditions of the proposition are *sufficient* conditions—these are not *necessary* conditions. If, contrary to our stipulation, Assumption 1 does not hold true while Assumption 2 is true, or even if both Assumption 1 and Assumption 2 do not hold true, then a ranking of the damage to o can

still be the same as the ranking of the damage to o'. The following example refers to a situation in which, due to an inherent dependency that the information system generates between errors and the corresponding correct values, the decomposability assumption is only approximately true (i.e., the product on the right hand side of (1) and (2) is approximately equal to the matching left hand side). GIGO, however, held true under most of the conditions that we have studied. Interestingly, our tests have indicated that the accuracy of a ranking based on the subsystem that we have identified was very high, approaching perfection. In conclusion, from a practical perspective, if two inputs that require ranking are processed within a small, identifiable subsystem, it may be worthwhile to study their ranking in that subsystem even when the validity of Assumption 1 and Assumption 2 may be compromised. Future work will continue to explore this direction.

ILLUSTRATION: APPLICATION OF THE RANKING STRATEGY

In essence, the proposed ranking strategy directs us to identify a minimal component of s that processes the inputs of interest, and center the ranking efforts on that component, s'. As long as the decomposability requirement and GIGO hold true, then a ranking of two inputs according to the damage to the output of s' is assured to be equal to their ranking in terms of the damage to the output of s.

In this section we illustrate the application of our damage ranking theory through the example of a popular class of applications which consist of multi-criteria, conjunctive or disjunctive decision or judgment rules. Instances of such applications include databases, expert systems that utilize domain knowledge in the form of multi-criteria satisficing decision rules, and classification models that use decision rules for their purpose.

Consider the following simple scenario, which centers on an organizational operational decision regarding a costly maintenance activity. Suppose that, in order to determine if a machine should or should not undergo this maintenance activity, decision makers employ a conjunctive decision rule which designates four decision variables: (1) the age of the machine, (2) its manufacturer, (3) its utilization status, and (4) its location. In particular, the maintenance decision applies the following criteria: (a) the machine is at least two years old *and* (b) the machine has been manufactured by ABC *and* (c) the machine is highly utilized (by some measure) *and* (d) the machine is kept in the east coast facility of the company in Massachusetts. As is often the case, the capital inventory database that serves this decision is not free of errors. Inaccurate age and manufacturer data are due to data entry errors or deficient communication among different organizational departments. Errors in data about machine utilization are mainly caused by delays in updating the data subsequent to utilization status changes. Errors in location data are often caused by delays in updating the data subsequent to equipment transfers from one facility to another. Obviously, errors in the data lead to errors in the classification of machines as passing, or not passing, the maintenance activity criteria. Assuming that the decision rule has been well chosen, both false negative and false positive decisions would be costly. False positive decisions should be avoided because of the high maintenance activity cost, while false negative decisions can lead to higher expected losses due to a higher rate of machine failures. Since errors in different decision inputs have, for the most part, different sources, an investment in the accuracy of one decision input would be largely separate from an investment in the accuracy of another decision input. To the degree that the decision makers have influence over the accuracy of the data, they can benefit from a tool that would rank the relevant attributes (manufacturing year, manufacturer, utilization status, and location) according to the damage that errors in each attribute inflict on the accuracy of the maintenance decision. Since resources (financial, human, etc.) are limited, a tool that assists in the identification of the inputs that would yield the highest gain in

decision accuracy can be useful.

In this scenario, the system s comprises the conjunctive rule that combines all four decision criteria. We assume that s is error-free (i.e., no processing errors). Suppose, for instance, that we want to rank two inputs of this decision rule, namely, the manufacturer of the machine (denoted next by v_i) and its location (denoted next by v_j) according to their damage values. Our strategy advises us to find a minimal subsystem s' that processes v_i and v_j and satisfies Assumption 1 (decomposability) and Assumption 2 (GIGO). A natural candidate for s' in this scenario is the binary conjunction operation that combines v_i and v_j. In other words, instead of accounting for the entire decision rule s we aim to limit the ranking effort to s', which designates the following binary conjunctive rule: the machine was manufactured by ABC *and* the machine is kept in Massachusetts.

We have developed analytical models that assist in the ranking of two inputs of a binary conjunctive rule and a binary disjunctive rule (Askira Gelman, 2010). Next, we will briefly present such a model that handles a conjunctive rule, and adapt it to our running example. Later we will examine the validity of Assumption 1 and Assumption 2.

Ranking Model

Let T_i and T_j denote the outputs of testing each of v_i and v_j, respectively, against the matching decision criterion. The possible outcomes of such a test are "false" (zero) and "true" (one). Specifically, v_i is tested against "ABC" to derive the value of T_i (e.g., if v_i=ACM then T_i=0, and if v_i=ABC then T_i=1) and v_j is tested against "MA" to derive the value of T_j. The values of T_i and T_j that are determined in this way are combined through a conjunction operation to generate the output of this component. Such an output, labeled o' in agreement with our earlier convention, can be either zero ("false" or "reject") or one ("true" or "accept").

Notation	Explanation
T_i , T_j	the outputs of testing each of v_i and v_j, respectively, against the matching decision criterion
E_i^v , E_j^v , E_i^T , E_j^T , $E^{O'}$, E^O	dichotomous variables that inform about error occurrence; the superscripts and subscripts describe the relevant variables
a_i^T , a_j^T	accuracy; the superscripts and subscripts describe the relevant variables
p_i^T , p_j^T	the probability that a given value satisfies the decision criterion on v_i , v_j, respectively
n_i, n_j	the number of different values that v_i , v_j, respectively, accepts
m_i , m_j	the number of different values of v_i , v_j, respectively, that meet the decision criterion

Table 2: New symbols in use by the illustration

While the above notation designates the correct inputs and outputs, an error in the recorded value is denoted by the letter E and a suitable superscript and/or subscript, e.g., E_i^v (the error in v_i), E_i^T (the error in T_i), $E^{O'}$, and so on. The variables E_i^T, E_j^T, $E^{O'}$, and E^O are dichotomous variables, since their values inform us about the *occurrence* of an error in the recorded values of T_i, T_j, o', and o (all of which are dichotomous variables themselves), respectively. A value of zero signifies that the recorded value of the variable is correct, while a value of one signifies that the value is incorrect. For instance, E_i^T =0 if the outcome of a test of the manufacturer is correct, and E_i^T =1 if the outcome of that test is incorrect (i.e., it is "false" when it should be "true" or vice versa). In contrast, E_i^v and E_j^v are not, in general, dichotomous. Since the data types of v_i and v_j can generally vary, a value of E_i^v or E_j^v may describe the occurrence of an error in a way that is similar to the former variables, or it may portray the error in a different way (e.g., show the magnitude of the error). In our scenario, however, since v_i and v_j are both categorical variables, E_i^v and E_j^v are dichotomous variables similar to the other error variables.

Accuracy (*measure*): The accuracy of the computed value of o', $a^{O'}$, is measured by the *probability of decision error*. The accuracy of the recorded value of T_i (T_j), denoted by a_i^T (a_j^T), is measured, again, by the *probability of error occurrence*. Since v_i and v_j are categorical variables, a_i^v and a_j^v are similarly measured by the *probability of error occurrence*.

Damage (*measure*): Our ranking model uses a partial derivative to implement the concept of damage. A derivative is a measure of the change in the output of a function when its input changes. A partial derivative is the derivative of a function of multiple variables when all but one variable of interest are held fixed. Therefore, it is consistent with the definition of damage.

The damage of the errors in v_i to the accuracy of the observed value of o' is calculated using (3):

$$\frac{\partial a^{O'}}{\partial a_i^v} = \{\Pr(T_j = 1 \mid E_i^T = 1) + \Pr(E_j^T = 1 \mid E_i^T = 1) - 2\Pr(T_i \cdot E_j^T = 1 \mid E_i^T = 1) - 2\Pr(T_j \cdot E_j^T = 1 \mid E_i^T = 1) + 2\Pr(T_i \cdot T_j \cdot E_j^T = 1 \mid E_i^T = 1)\} \cdot \frac{\partial a_i^T}{\partial a_i^v} \quad (3)$$

According to (3), the damage of errors in v_i to o' is a function of several parameters. Mainly, we note that the damage reflects the probabilities that the decision criteria are met on error-free inputs, the error probabilities when testing for these criteria, and also any statistical interdependencies among such events. In other words, it reflects characteristics of the error-free inputs, the process, and error characteristics, and also accounts for interdependencies. For instance, the term $\Pr(T_j = 1 \mid E_i^T = 1)$ is determined by the probability of error occurrence when testing for the decision criterion on v_i, and the probability of its joint occurrence with the condition that the decision criterion on v_j is satisfied. In the absence of statistical dependencies

between these events, this term is simply the probability that the decision criterion on v_j is satisfied.

The validity of (3), which applies to any two inputs that are combined by a conjunctive rule, follows directly from (Askira Gelman, 2010). (A formulation of the damage of errors in v_j is comparable to (3), given appropriate notation adjustment.)

Now, suppose that the variables in $\{v_i, E_i^v, v_j, E_j^v\}$ are statistically independent. Suppose also that v_i accepts n_i different values (n_i is the number of different equipment manufacturers) and m_i of these values match the decision criterion. Since one manufacturer, "ABC," is of special interest, m_i=1. Similarly, suppose that v_j accepts n_j different values, m_j of which satisfy the decision criterion. Again, m_j=1 in our scenario. Let p_i^T (p_j^T) denote the probability that a given value satisfies the decision criterion on v_i (v_j). Suppose that errors in the recorded values of v_i and v_j are uniformly distributed, such that the probability of incorrectly showing any of the other n_i-1 (n_j-1) values is $a_i^v/(n_i-1)$ ($a_j^v/(n_j-1)$) (assume that the system prevents us from entering values that are not included in the lists of recognized values). Under these assumptions, the damage formulation (3) can be re-written in terms of p_i^T, p_j^T, n_i, n_j, a_j^v, m_i, and m_j:

$$\frac{\partial a_i^{O'}}{\partial a_i^v} = p_j^T(m_i + p_i^T \cdot n_i - 2p_i^T \cdot m_i)/(n_i - 1) + a_j^v/((n_j-1)(n_i-1)) \cdot \{(m_j + p_j^T \cdot n_j - 2p_j^T \cdot m_j) \cdot (m_i - p_i^T \cdot n_i) - 2 \cdot p_j^T \cdot (n_j - m_j) \cdot (m_i - p_i^T \cdot m_i)\} \quad (4)$$

Equation (4) shows that the damage depends on input accuracy, and in addition, it accounts for characteristics of the correct data and the nature of the process. In particular, it factors in the number of different values that each input accepts, the number of different values that satisfy the decision criterion, and, similarly, the probability that a given value meets the criterion.

A proof of (4) is shown in the Appendix. Implementation of this damage ranking model is shortly discussed by (Askira Gelman, 2010); see also (Askira Gelman, forthcoming). Once the damage values for v_i and v_j are known, they can be compared to determine the damage ranking.

We now turn to the question of the validity of Assumption 1 and Assumption 2.

Assumption 1 (Decomposability): When s' is a single binary conjunctive (or disjunctive) operation and s is a multi-criteria rule such that s' is a component of s, then the decomposability assumption is not satisfied in full. The right hand side of (1) and (2) is only approximately equal to the matching left hand side. This deviation is rooted in a statistical dependency that conjunction and disjunction operations generate between errors and the corresponding correct values (Askira Gelman, 2009b). For the purpose of the present paper we have conducted a numerical analysis in order to directly explore the deviation of equations (1) and (2) from the actual damage value. We studied conjunctive rules with up to five decision variables. In order to simplify the analysis, we centered on T_i, T_j, E_i^T, and E_j^T, which are statistically independent of each other, much like the corresponding parameters that describe the other inputs of the conjunctive rule. Our analysis accounted for values of p_i^T, p_j^T, and respective parameters of the other decision inputs which vary anywhere in the range 0.01-0.99. The values of a_i^T, a_j^T, and

respective parameters of the other inputs varied in the range 0.01-0.10 (these values can match much higher fundamental error rates, i.e., a_i^v and a_j^v can be substantially higher). The results showed that in over 70% of the cases the deviation of the actual damage values from the damage values calculated from (1)-(2) was less than 15%.

Assumption 2 (GIGO): Earlier work has shown, both analytically and empirically, that, due to the nature of the conjunction and disjunction operations, GIGO does not always hold true in satisficing decision rules (Askira Gelman, 2007). That work has also characterized conditions in which GIGO is violated. Nonetheless, here we provide additional information on these conditions. Such information has been obtained through a series of Monte Carlo simulations that we have conducted as part of this study. These simulations examined the output of conjunction of two inputs. The simulations ignored the values of the decision variables v_i and v_j since our earlier research had found such variables largely irrelevant for the phenomenon of interest. Instead, the simulation focused directly on T_i and T_j and the corresponding error terms, E_i^T and E_j^T. The values of these four variables were generated such that they assumed statistical independence, but covered a wide range of probabilities. The values of p_i^T and p_j^T varied from 0.01 to 0.99 in increments of 0.01, and the values of a_i^T and a_j^T varied from 0.01 to 0.50 in increments of 0.01 (i.e., one simulation for each possible combination of the listed values of p_i^T, p_j^T, a_i^T, and a_j^T). In essence, the chosen probability combinations established a sample of nearly all practical probability combinations. In order to determine the validity of GIGO, each simulation increased the value of each of a_i^T and a_j^T, in turn, by 0.01, and compared the subsequent decision error rates to the base rate. The results of the simulations show a violation of GIGO in roughly 9% of the simulations. However, just about 1.5% of the simulations in which a_i^T and a_j^T were limited to a maximum of 0.10 violated GIGO. From a practical perspective, the latter result is probably more meaningful than the result that shows a 9% deviation from GIGO. The reason is, again, that the upper boundary on a_i^T and a_j^T typically corresponds to a much higher boundary on the fundamental input error rates, a_i^v and a_j^v. Therefore, a boundary of 0.1 probably covers realistic data error rates.

In conclusion, although Assumption 1 and Assumption 2 are not fully satisfied in this scenario, these assumptions seem to offer useful approximations of the actual conditions. Our research has demonstrated that the damage ranking model (3) provided highly accurate predictions of the actual damage ranking, up to 99% of the tested instances (Askira Gelman, 2009). A major advantage of this model is that, regardless of the size of the decision rule, the model employs parameters of two inputs only. Hence, this model simplifies the damage ranking task.

CONCLUDING REMARKS

In this paper we propose a broad strategy for ranking the inputs of an information system according to the damage that errors in each input inflict on the output of the information system. Two important elements of this strategy are the notion of damage and the associated idea of employing damage ranking in data management decision making. A major potential advantage of this strategy is its relative simplicity.

Future studies should continue to explore the practical usefulness of this approach. The value of this strategy should be empirically assessed through feasibility and implementation studies. Can small subcomponents such as we have delineated be easily identified? To what extent are

Assumption 1 and Assumption 2 valid? How critical are they for the proposed strategy? What is the amount of effort that is required in order to produce the ranking given that Assumption 1 and Assumption 2 can be taken to be valid in a chosen setting? These questions will have to be addressed through future work. Obviously, although the illustrative example that we have presented included an analytical model of the damage, such a model is not a necessary condition for this approach to succeed in general. An assessment of the damage can be carried out in various ways. The benefit that this work promises, i.e., that the ranking task will be simplified through a focus on a smaller subsystem, can be obtained regardless of the former choice.

REFERENCES

[1]. Askira Gelman, I., Simulations of the Relationship between an Information System's Input Accuracy and its Output Accuracy. *9th International Conference on Information Quality (ICIQ)*, MIT, Cambridge MA, 2004.

[2]. Askira Gelman, I., GIGO or not GIGO: Error Propagation in Basic Information Processing Operations. *13th Americas Conference on Information Systems (AMCIS)*, 2007.

[3]. Askira Gelman, I., Simulations of Error Propagation for Prioritizing Data Accuracy Improvements in Multi-Criteria Satisficing Decision Making Scenarios. *30th International Conference on Information Systems (ICIS)*. Phoenix, Arizona, 2009.

[4]. Askira Gelman, I., The Asymmetric Nature of Decision Errors in Multi-Criteria, Satisficing Decisions. *15th Americas Conference on Information Systems (AMCIS)*, 2009b.

[5]. Askira Gelman, I.**,** *Setting Priorities for Data Accuracy Improvements in Satisficing Decision-making Scenarios: A Guiding Theory*. *Decision Support Systems (DSS)*, Vol. 48, No. 4, 2010, pp. 507-520.

[6]. Askira Gelman, I., A Model of Error Propagation in Conjunctive Decisions and its Application to Database Quality Management. *Journal of Database Management (JDM)*. Forthcoming.

[7]. A. Avenali, C. Batini, P. Bertolazzi, And P. Missier, Brokering Infrastructure For Minimum Cost Data Procurement Based On Quality–Quantity Models, *Decision Support Systems* 45(1) (2008).

[8]. Ballou, D. P. And Pazer, H. L. (1985) Modeling Data And Process Quality In Multi-Input, Multi-Output Information Systems. *Management Science*, Vol. 31, No. 2, Pp. 150-162

[9]. D. P. Ballou And G.K. Tayi, Methodology For Allocating Resources For Data Quality Enhancement, Communications Of The Acm 32(3) (1989).

[10]. Ballou, D.P., And Pazer, H.L. (1990). A Framework For The Analysis Of Error In Conjunctive, Multi-Criteria, Satisficing Decision Processes. *Decision Sciences*, 21(4), Pp. 752-770.

[11]. Ballou, D. P., Wang, R. Y., Pazer, H. L., And Tayi, G. K. (1998) Modeling Information Manufacturing Systems To Determine Information Product Quality. *Management Science*, Vol. 44, No. 4, Pp. 462-484.

[12]. Eckerson, Wayne W. (2002) Achieving Business Success Through A Commitment To High Quality Data. *Tdwi Report Series, The Data Warehousing Institute*.

[13]. Even, A., And Shankaranarayanan, G. (2007) Utility-Driven Assessment Of Data Quality. *The Data Base For Advances In Information Systems*, Vol. 38, No. 2, Pp. 75-93.

[14]. Fisher C. W., Lauria, E.J.M., and Matheus, C.C. An Accuracy Metric: Percentages, Randomness, and Probabilities. *ACM Journal of Data and Information Quality (JDIQ)*, Vol. 1, No. 3, 2009.

[15]. Hemilton, M., and Zeldin, S. Higher order software-A methodology for defining software. *IEEE Transactions on Software Engineering*, Vol. SE-2, No. 1, pp. 9-32, 1978.

[16]. Hevner, S. March, J. Park, And S. Ram, "Design Science Research In Information Systems," *Management Information Systems Quarterly* (28:1), March 2004, Pp. 75-105.

[17]. Janson, M. (1988) Data Quality: The Achilles Heel Of End-User Computing, *Omega: International Journal Of Management Science* 16(5).

[18]. Juran, J.M. (1988) *Juran On Planning For Quality* (The Free Press, New York).

[19]. Lee, Y. Strong, D. Kahn, B. And Wang, R. (2002) Aimq: A Methodology For Information Quality Assessment, *Information And Management* 40(2).

[20]. March, S. And Smith, G. "Design And Natural Science Research On Information Technology," *Decision Support Systems* 15, 1995, Pp. 251 - 266.

[21]. Motro, A., And Rakov, I. Not All Answers Are Equally Good: Estimating The Quality Of Database Answers. In *Flexible Query-Answering Systems* (T. Andreasen, H. Christiansen, And H.L. Larsen, Editors). Kluwer Academic Publishers, 1997, Pp. 1-21.

[22]. Naumann, F., Leser, U., And Freytag, J. (1999) Quality-Driven Integration Of Heterogeneous Information Systems. In *Proceedings Of The 25th International Conference On Very Large Data Bases*, Vldb 99.

[23]. Parsaye K. And Chignell, M. (1993) Data Quality Control With Smart Databases, *Ai Expert* 8(5).

[24]. Parssian. A., Sarkar, S., And Varghese, S.J. (2004) Assessing Data Quality For Information Products: Impact Of Selection, Projection, And Cartesian Product. *Management Science*, Vol. 50, No. 7, Pp. 967-982.

[25]. Parssian. A., (2006) Managerial Decision Support With Knowledge Of Accuracy And Completeness Of The Relational Aggregate Functions. *Decision Support Systems*, Vol. 42, Pp. 1494-1502.

[26]. Pierce, B. (2002). *Types and Programming Languages*. MIT Press.

[27]. Redman, T.C. "Data: An Unfolding Disaster," *Dm Review Magazine*, 2004.

[28]. Shankaranarayan, G., Zaid M., And Wang, R. (2003) Managing Data Quality In Dynamic Decision Environments: An Information Product Approach. *Journal Of Database Management*, Vol. 14, No. 4.

[29]. Wand, Y. and Weber, R., An Onthological Model of an Information System, IEEE Transactions on Software Engineering, Vol. 16, No. 11, 1990.

[30]. Wang, R., Reddy, M., And Kon, H. (1995) Toward Quality Data: An Attribute-Based Approach, *Journal Of Decision Support Systems*, Vol. 13, No. 3-4., Pp. 349-372.

[31]. Wang, R.Y., And Strong, D.M. (1996) Beyond Accuracy: What Data Quality Means To Data Consumer. *Journal Of Management Information Systems*, 12, Pp. 5-34.

APPENDIX

Proof of Equation (4):

1. The probability that a value of v_j will incorrectly fail the decision criterion is given by
$$\Pr(T_j \cdot E_j^T = 1) = p_j^T \cdot a_j^v \cdot (n_j - m_j)/(n_j - 1).$$
2. Under the assumed conditions we see that:
$$a_j^T = a_j^v \cdot \{(1 - p_j^T) \cdot m_j + p_j^T \cdot (n_j - m_j)\}/(n_j - 1) = a_j^v \cdot (m_j + p_j^T \cdot n_j - 2p_j^T m_j)/(n_j - 1)$$
3. The conditional probability that the value of v_i satisfies the decision criterion given a classification error, is :
$$\Pr(T_i = 1 \mid E_i^T = 1) = p_i^T \cdot (n_i - m_i)/(m_i + p_i^T \cdot n_i - 2p_i^T m_i)$$
4. Similar to a_j^T, we see that a_i^T is given by:
$a_i^T = a_i^v \cdot \{(1 - p_i^T) \cdot m_i + p_i^T \cdot (n_i - m_i)\}/(n_i - 1) = a_i^v \cdot (m_i + p_i^T \cdot n_i - 2p_i^T m_i)/(n_i - 1)$. Therefore,
$$\frac{\partial a_i^T}{\partial a_i^v} = (m_i + p_i^T n_i - 2p_i^T m_i)/(n_i - 1)$$
5. Equation (3) states that:
$$\frac{\partial ao'}{\partial a_i^v} = \{\Pr(T_j = 1 \mid E_i^T = 1) + \Pr(E_j^T = 1 \mid E_i^T = 1) - 2\Pr(T_i \cdot E_j^T = 1 \mid E_i^T = 1) - 2\Pr(T_j \cdot E_j^T = 1 \mid E_i^T = 1)$$
$$+2\Pr(T_i \cdot T_j \cdot E_j^T = 1 \mid E_i^T = 1)\} \cdot \frac{\partial a_i^T}{\partial a_i^v}$$

Therefore, under our independence assumptions:
$$\frac{\partial a_i^{O'}}{\partial a_i^v} = \{p_j^T + a_j^T - 2a_j^T \cdot \Pr(T_i = 1 \mid E_i^T = 1) - 2\Pr(T_j \cdot E_j^T = 1) + 2\Pr(T_i = 1 \mid E_i^T = 1) \cdot \Pr(T_j \cdot E_j^T = 1)\} \cdot \frac{\partial a_i^T}{\partial a_i^v}$$

We now insert (1)-(4) in (5), and simplify, to get:
$$\frac{\partial a_i^{O'}}{\partial a_i^v} = [p_j^T + \{a_j^v/(n_j - 1)\} \cdot \{(m_j + p_j^T \cdot n_j - 2p_j^T \cdot m_j) \cdot (1 - 2 \cdot p_i^T \cdot (n_i - m_i)/(m_i + p_i^T \cdot n_i - 2p_i^T \cdot m_i))$$
$$-2 \cdot p_j^T \cdot (n_j - m_j)(1 - p_i^T \cdot (n_i - m_i)/(p_i^T \cdot n_i + m_i - 2p_i^T \cdot m_i))\}] \cdot (m_i + p_i^T \cdot n_i - 2p_i^T \cdot m_i)/(n_i - 1) =$$
$$p_j^T (m_i + p_i^T \cdot n_i - 2p_i^T \cdot m_i)/(n_i - 1) + a_j^v/((n_j - 1)(n_i - 1)) \cdot \{(m_j + p_j^T \cdot n_j - 2p_j^T \cdot m_j) \cdot (m_i - p_i^T \cdot n_i)$$
$$-2 \cdot p_j^T \cdot (n_j - m_j) \cdot (m_i - p_i^T \cdot m_i)\}$$
End of proof.