IMPARTIAL VERSUS UTILITY-DRIVEN ASSESSMENT OF DATA QUALITY: METHODOLOGY, INSIGHTS, AND IMPLICATIONS FOR MANAGING CUSTOMER DATA

A. Even[1], G. Shankaranarayanan[2]

[1]Department of Industrial Eng. and Mgmt., Ben-Gurion University of the Negev,
P.O. Box 653, Beer-Sheva 84105, Israel

[2]Information Systems Department, Boston University School of Management,
595 Commonwealth Ave., Boston, MA, 02215, USA

ABSTRACT

This study presents a methodology for dual assessments of data. Impartial assessment measures the extent to which data is defective. Utility-driven assessments of data quality measure the extent to which the presence of quality defects degrades data utility – the benefit gained from using that data in a specific business setting. The dual assessment methodology is demonstrated in a real-world setting using alumni data – a large data resource for managing alumni relations and initiating pledge campaigns. Dual-assessment results provide important inputs that can direct the implementation and management of quality improvement policies in this data resource.

Keywords: data quality, data utility

1. INTRODUCTION

High data quality is critical for successful integration of information systems, as the presence of data quality defects degrades usability and damages revenues and credibility [3]. However, with the rapidly increasing data volumes, high data quality is harder to achieve and sustain. Targeting defect-free data is expensive, often practically impossible and, from an economic perspective, might be sub-optimal, as the cost of improving quality may offset the benefits gained. Given these challenges, data quality management must define: *(a) Quality Targets:* the targeted level can be evaluated along a continuum: perfect data quality at one end (i.e., no defects), and a "hands off" approach at the other end (i.e., accepting data quality as is). Between these ends, we may consider improving quality to some extent while permitting

imperfections. *(b) Priorities:* we may consider an equal treatment of all records and attributes or, a differentiating policy – giving higher priority to improving the quality of certain records and/or attributes, and possibly making no significant efforts to improve others.

Given targets and priorities, different treatments and policies for improving data quality are: *(a) Prevention:* reducing defect rates during data acquisition and processing - e.g., by improving user interfaces, disallowing missing values, validating against a value domain, enforcing integrity, or using cleaner (and possibly more expensive) data sources. *(b) Auditing:* defects may also occur during data processing (e.g., due to miscalculations, or code-mismatch when integrating multiple sources), or after data has been stored (e.g., due to changes to the corresponding real-world entity). Overcoming such defects requires auditing records, monitoring the process, and detecting the existence of defects. *(c) Correction:* correcting defects is often time consuming and costly (e.g., when a customer has to be contacted, or when missing content has to be purchased). *(d) Usage:* one might recommend not using certain records and/or attributes, or prevent usage altogether – e.g., when the quality is too low and cannot be improved.

Quantitative data quality assessments can provide important inputs and direct quality improvement efforts. Today, such assessments are mostly impartial, measuring the extent to which quality defects exist, disregarding usage context. Research in data quality has highlighted the importance of contextual assessment [2], but does not minimize the value of impartial quality assessments. We suggest that quality assessment can be enhanced by considering data utility – a quantitative assessment of the benefits gained from data within specific context (e.g., a decision task). The same data may have different utility in different usage contexts and, accordingly, the presence of defects may differently impact utility degradation. We therefore suggest that measuring quality as the extent to which utility is degraded, affords a contextual assessment of the impact of quality defects. We develop a methodology that measures both impartial and utility-driven quality along different quality dimensions (illustrated here using completeness and currency). The results of this dual evaluation offer insights on quality characteristics and guide the development of quality improvement policies. We demonstrate this methodology in the context of Customer Relationship Management (CRM), using large samples from a data resource used by a large university for managing alumni relations. In the remainder of this paper, we first reveiw the analytical baseline for the dual methodology for quality assessment. We then demonstrate an

application of this methodology with alumni data and use the results to formulate quality improvement policies for this data resource. To conclude, we highlight the contributions of this study, discuss managerial implications, and propose directions for further research.

2. IMPARTIAL VERSUS UTILITY-DRIVEN ASSESSMENT OF DATA QUALITY

This study adopts the measurement framework suggested in [1]. This framework, briefly described here, permits contextual measurement of quality along different dimensions and, with certain relaxations, allows impartial assessment as well. Quality measurement in this framework is driven by utility - a non negative measurement of its value contribution. The evaluated dataset has N records (indexed by *[n]*), and M attributes (indexed by *[m]*). The data content of attribute *[m]* in record *[n]* is denoted $f_{n,m}$. The quality measure $q_{n,m}$ reflects the extent to which attribute *[m]* of record *[n]* suffers from a quality defect (between *0* - severe defects, and *1* - no defects). The overall utility U^D is attributed along records $\{U_n\}$, based on relative importance such that $U_D=\Sigma_{n=1..N}U_n$. The utility-mapping function used in this framework links record contents and quality to its utility:

(1) $U_n = u_i\left(\left\{f_{n,m}\right\}_{m=1..M}, \left\{q_{n,m}\right\}_{m=1..M}\right)$

For a given set of attribute contents $\{f_{n,m}\}$, record utility reaches an upper limit U^{MAX}_n when all attributes have perfect quality (i.e., $\{q_{n,m}=1\}$) and may be reduced by an extent when certain attributes are defective. The record quality Q_n is defined as a *[0,1]* ratio between the actual utility U_n and the upper limit U^{MAX}_n:

(2) $Q_n = U_n / U_n^{MAX} = \left(u\left(\left\{f_{n,m}\right\}_{m=1..M}, \left\{q_{n,m}\right\}_{m=1..M}\right)\right) / \left(u\left(\left\{f_{n,m}\right\}_{m=1..M}, \left\{q_{n,m}=1\right\}_{m=1..M}\right)\right)$

Similarly, dataset quality Q^D is the ratio between actual and maximum possible utility:

(3) $Q^D = \left(\sum_{n=1..N} U_n\right) / \left(\sum_{n=1..N} U_n^{MAX}\right) = \left(\sum_{n=1..N} U_n^{MAX} Q_n\right) / \left(\sum_{n=1..N} U_n^{MAX}\right)$

When utility is allocated independent of attribute content (i.e., constant $U^{MAX}_n=U^D/N$), the result is an impartial measure that reflects a ratio between the counts of perfect items and total items, which is consistent with common structural definitions (e.g., [2, 3]):

(4) $Q_n = (1/M)\sum_{m=1..M} q_{n,m}$, and $Q^D = (1/MN)\sum_{n=1..N}\sum_{m=1..M} q_{n,m}$

This definition permits measurement along different dimensions, each reflecting a specific type of quality defect. For example, completeness reflects missing or corrupted values, validity reflects failure to conform to a value-domain, accuracy reflects incorrect content, and currency reflects the extent to which data items are not up-to-date.

Not all records in a dataset contribute equally to utility. The likelihood of the occurrence of quality defects in a dataset record may be independent of its utility. However, recognizing a record as having a higher utility may encourage more focused efforts to reduce quality defects in it. Utility-driven measurement reflects the impact of quality defects on the value contribution of the data, i.e., the extent to which utility is reduced by the presence of defects. Comparing the results of utility-driven assessments to impartial assessments is important for managing data quality in such datasets. At a high-level, we can differentiate between three cases with respect to such a comparison in large datasets: *(a) Utility-driven scores are significantly higher than impartial scores:* this indicates that records with high utility are less defective. Two possible explanations are: first, defective records are less usable to begin with, hence, have inherently lower utility. Second, some differentiating error-correction policies may have been applied – some efforts were made to maintain records with higher utility at a high quality level and eliminate their defects. *(b) Utility-driven scores not significantly different from impartial scores:* this indicates no association – the proportion of quality defects does not depend on the utility of certain records, whether high or low. This may also indicate high equality – utility that is nearly evenly distributed between all records, and *(c) Utility-driven scores significantly lower than impartial scores:* this indicates that records with high utility have a higher rate of quality defects. This abnormality may indicate a systematic cause of defects for record with high utility. This may also indicate high inequality in the dataset (i.e., a large proportion of utility associated with a small number of records), and some substantial damage to high-utility records. Understanding the relationships between impartial measurement and utility-driven measurement can help develop DQM policies, as demonstrated with our empirical assessment of the alumni data.

3. ASSESSING THE QUALITY OF ALUMNI DATA

To demonstrate utility-driven assessment of quality and its implications for prioritizing quality improvement efforts, we evaluate a sizably large sample of alumni data. This critical data resource helps generate a significant portion of the university's revenue. The alumni data is used by different departments for contacting donors, tracking their gift history and managing pledge campaigns. This data resource, and the system that manages it, can be viewed as a form of Customer Relationship Management (CRM). Such systems are used for

managing donor relations, tracking their past contributions, analyzing gifting patterns, and segmenting them for better targeting future promotion campaigns.

We examine dual assessment of data quality with samples from two datasets in the alumni database managed by a large university. We consider two key datasets: *(a) Profiles (358,372 records)* - Contact and demographic data on alumni and other potential donors, including attributes such as Profile ID, Graduation Year, School of Graduation, Gender, Marital Status, Income, Ethnicity, and Religion. *(b) Gifts (1,415,432 records)* - The history of donations made, with attributes such as Gift ID, Profile ID (a foreign key to the Profiles dataset), Gift Year (derived from the gift date), and the Gift Amount. The data has been collected between 1983 and 2006. In 1983 and 1984 (soon after system initiation), a bulk of records that reflect prior activity were added (203,359 profiles, 405,969 gifts), and since then both datasets have grown gradually. To evaluate inequality within both profile and gift records, we use the gifts made in 2006 as a proxy for utility. To preserve confidentiality, we multiplied the gift amounts shown in this study by a constant. The results of our evalution are summarized in the following table, and further described in the following paragraphs.

		Impartial Quality	**Utility-Driven Quality**			
			Inclination (1 Year)	**Inclination (2-5 Years)**	**Amount (1 Year)**	**Amount (2-5 Years)**
Attributes Completeness	*School*	0.999	0.999	0.999	0.999	0.999
	Gender	0.990	0.997	0.998	0.997	0.999
	Marital	0.894	0.950	0.958	0.984	0.977
	Income	0.631	0.872	0.896	0.891	0.836
	Ethnicity	0.596	0.646	0.654	0.656	0.496
	Religion	0.605	0.717	0.715	0.819	0.751
	All	0.786	0.863	0.870	0.891	0.843
Record Completeness	*Absolute*	0.356	0.497	0.511	0.561	0.608
	Grade	0.786	0.863	0.870	0.891	0.843
Record Currency	*Recent-1*	0.171	0.282	0.219	0.635	0.552
	Recent-5	0.510	0.635	0.635	0.899	0.860

First, we computed the following variables for each record:

a) *Missing-Value Indicators:* for each attribute (6 overall), the corresponding variable reflects whether the value is missing (=0) or not (=1). We also computed the absolute rank (0 if at least one attribute is missing, 1 otherwise), and the grade rank (the average of the 6 attribute indicator), for each record.

b) *b) Up-to-date:* we calculate two binary currency indicators – one indicating whether a record has been updated within the last 1-year period, and the other indicating update in the last 5-year period.

c) *c) Utility Measurements:* We have computed the inclination to donate (0 or 1) and the total donation amount, each for the last 1 year (2006) and the previous 4 years (2002-2005).

Impartial quality score use the ratio measurements based on item-counts (Equation 4), while for utility-driven quality assessment we apply the weighted-average formulation (Equation 3), using the four utility measures as weights. Notably most utility-driven data quality measurement scores are higher than their corresponding impartial measurement scores. This is not surprising since, along most indicators, higher utility has a significantly stronger association with higher impartial quality. However, some insights can be gained by observing the extent to which utility-driven measurements are higher and more consistent:

- Utility-driven completeness measurements, at the attribute level and at the record level, are consistent along the four utility metrics. This implies that, when assessing the completeness of this alumni profile data, calculating utility-driven measurements along multiple utility metrics does not grant a significant advantage over measuring it along a single metric
- For attributes with inherently high impartial completeness (e.g., School and Gender), utility-driven measurements are not substantially different from the impartial measurement scores. Some margin exists for Marital Status – but since the impartial completeness is relatively high, this margin is fairly small.
- For attributes with inherently low impartial quality, we see substantial differences in the margin between the impartial and the utility-driven scores. Considering Ethnicity, the margin is relatively minor. It is slightly higher for Religion, and a lot higher for Income. This implies that these attributes have very different association with the utility gained. The completeness of Income attribute significantly differentiates between low-utility and high-utility profile records (both along Inclination and Amount). The completeness of Religion data also

differentiates these, but to a lesser extent, and the completeness of Ethnicity does not significantly differentiate the profile records.

- Measuring completeness for all the attributes combined, or measuring it at the record level, has an averaging effect. Some margins exist between impartial and utility-driven scores, but they are not as significant as the margins for the measurements of specific attributes.
- Unlike completeness, with respect to currency, amount-driven scores are significantly higher than inclination-driven scores along all indicators. This implies that the extent to which a record is up-to-date is significantly associated with the amount donated, beyond just the fact that a person has made a donation. This finding may suggest that the current practice is to audit and update more frequently data on donors who have contributed or have a high contribution potential (as confirmed by the alumni data administrators).
- With respect to utility-driven currency measurement, there is a significant difference between using the inclination versus using amount as utility factors. However, there is no significant difference between evaluating utility (using inclination or amount as factors) over 1 year versus the previous 4 years.

The results of our evaluation shed light on a few issues that need further attention and can guide the development of better quality management policies:

Differentiation: In general, the data administrators should clearly consider a differentiating policy with respect to auditing records and attributes, correcting quality defects, and implementing procedures to prevent defects from reoccurring. They may also consider recommending that data users refrain from using certain records or attributes for certain types of usages (decision tasks and applications).

Attributing Utility: Our results highlight the benefit of measuring and attributing utility. Our metrics, inclination and amount, reflect the impact of quality defects on utility; hence, permit convenient calculation of utility-driven measurements.

Improving Completeness: The results indicate that analyzing the impact of missing values at the record level alone is insufficient. There is certainly a need to further assess the impact of missing values at the attribute level. The impartial completeness of certain attributes is inherently high (e.g., School and Gender, with nearly 0 missing values); hence, the potential to gain utility by correcting these attributes is negligible. Even for attributes with lower impartial completeness, we can expect substantial variability – with some attribute (e.g., Income) we see a strong association between missing values and utility contribution. Such

attributes obviously need to get a very high priority in terms of improvement efforts. With other attributes (e.g., Marital Status and Religion), we see some association, but to a lesser extent. With yet other attributes (e.g., Ethnicity), the association, if at all, is very small. In the latter case, we may reconsider whether it is worthwhile to invest in any quality improvement efforts, or even consider giving up the storage and management of this attribute. Notably, the data resource evaluated here contains many (over a hundred) other profile attributes, and managing these could benefit from a similar evaluation.

Improving Currency: Utility was strongly linked to currency – outdated profiles are associated with lower inclination and amount. This indicates a need to audit profiles more often. As shown earlier, there is a strong association between recent donations (last 1 year) and past donations (previous 4 years); hence, profiles that are associated with recent inclination should get high priority for quality improvement (e.g., a more frequent auditing).

The quality and the utility of alumni data can certainly be improved, as only a relatively small number of profiles are associated with donations, and quality defects are present in high proportions. Importantly, our analyses do not offer a comprehensive solution for prioritization and policies, but rather demonstrates the methodology and the concrete insights to be gained from such analyses. A more complete solution demands an analysis of all relevant attributes, evaluation of other utility measurements, statistical tools for estimation of future benefits, and possibly a revision of existing data usage patterns.

4. CONCLUSIONS

Quantitative quality assessment is important for continuous improvement of data quality. Common measurement methods tend to reflect an impartial perspective and disregard the context in which the data is used. This study explores a new measurement methodology that reflects a contextual perspective as well, by observing not only the presence of defects, but also their impact on the utility gained. Applying both impartial and utility-driven assessments provides important insights on strengths and weaknesses with current data quality management practices. It can direct the improvement of these practices and the development of new policies. The application of this methodology is demonstrated in the context of managing alumni data, showing how current quality measurement methods compare, and are supplemented by, the proposed method for measuring and improving data quality.

The results highlight the importance of assessing the utility of data resources. Different elements in a dataset (e.g., records and/or attributes) may significantly vary in their contribution to utility. Modeling and quantifying utility distribution and detecting possible inequalities can direct quality improvement efforts and help prioritize them. Utility assessment is also important in the presence of significant economic tradeoffs – certain improvement efforts are expensive, and their cost might offset the added utility. Evaluating both utility and cost help assess these tradeoffs and detect economically-optimal policies.

BIBLIOGRAPHY

Even, A., and Shankaranarayanan, G. “Assessing Data Quality: a Value-Driven Approach”, The DATA BASE (38:2), May 2007, pp. 76-93

Pipino L.L, Yang, W.L. and Wang, R.Y. “Data Quality Assessment,” Communications of the ACM (45:4), April 2002, pp 211-218

Redman, T.C. Data Quality for the Information Age, Artech House, Boston, MA, 1996