INTRODUCTION

For many years the idea that for high information processing systems effectiveness, high quality of data is not less important than the systems’ technological perfection was not widely understood and accepted. The way to understanding the complexity of the data quality notion was also long, as will be shown in this paper. However, progress in modern information processing systems development is not possible without improvement of data quality assessment and control methods. Data quality is closely connected both with data form and value of information carried by the data. High-quality data can be understood as data having an appropriate form and containing valuable information. Therefore, at least two aspects of data are reflected in this notion: (1) technical facility of data processing and (2) usefulness of information supplied by the data in education, science, decision making, etc.

BACKGROUND

In the early years of information theory development a difference between the quantity and the value of information was noticed; however, originally little attention was paid to the information value problem. Hartley (1928), interpreting information value as its psychological aspect, stated that it is desirable to eliminate any additional psychological factors and to establish an information measure based on purely physical terms only. Shannon and Weaver (1949) created a mathematical communication theory based on statistical concepts, fully neglecting the information value aspects. Brillouin (1956) tried to establish a relationship between the quantity and the value of information, stating that for an information user, the relative information value is smaller than or equal to the absolute information (i.e., to its quantity, chap. 20.6). Bongard (1960) and Kharkevitsch (1960) have proposed to combine the information value concept with the one of a statistical decision risk. This concept has also been developed by Stratonovitsch (1975, chaps. 9-10). This approach leads to an economic point of view on information value as profits earned due to information using (Beynon-Davies, 1998, chap. 34.5). Such an approach to information value assessment is limited to the cases in which economic profits can be quantitatively evaluated. In physical and technical measurements, data accuracy (described by a mean square error or by a confidence interval length) is used as the main data quality descriptor (Piotrowski, 1992). In medical diagnosis, data actuality, relevance and credibility play a relatively higher role than data accuracy (Wulff, 1981, chap. 2). This indicates that, in general, no universal set of data quality descriptors exists; they rather should be chosen according to the application area specificity. In the last years data quality became one of the main problems posed by World Wide Web (WWW) development (Baeza-Yates & Ribeiro-Neto, 1999, chap. 13.2). The focus in the domain of finding information in the WWW increasingly shifts from merely locating relevant information to differentiating high-quality from low-quality information (Oberweis & Perc, 2000, pp. 14-15). In the recommendations for databases of the Committee for Data in Science and Technology (CODATA), several different quality types of data are distinguished: (1) primary (rough) data, whose quality is subjected to individually or locally accepted rules or constraints, (2) qualified data, broadly accessible and satisfying national or international (ISO) standards in the given application domain, and (3) recommended data, the highest quality broadly accessible data (like physical fundamental constants) that have passed a set of special data quality tests (Mohr & Taylor, 2000, p. 137).

BASIC PROBLEMS OF DATA QUALITY ASSESSMENT

Taking into account the difficulty of data value representation by a single numerical parameter Kulikowski (1969) proposed to characterize it by a vector whose components describe various easily evaluated data quality aspects (pp. 89-105). For a multi-aspect data quality evaluation, quality factors like data actuality, relevance, credibility, accuracy, operability, etc. were proposed. In the ensuing years the list of proposed data quality factors by other authors has been extended up to almost 200 (Pipino, Lee, & Wang, 2002; Shanks & Darke, 1998; Wang & Storey, 1995; Wang Strong, & Firth, 1996). However, in order to make data quality assessment possible it is not enough to define a large number of sophisticated data quality factors. It is also necessary to establish the meth-
Data Quality Assessment

ods of multi-aspect data quality comparison. For example, there may arise a problem of practical importance—What kind of data in a given situation should be preferred: the more fresh but less credible, or outdated but very accurate ones. A comparison of multi-aspect data qualities is possible when the following formal conditions are fulfilled: (1) The quality factors are represented by non-negative real numbers, (2) if taken together, they form a vector in a linear semi-ordered vector space, and (3) they are defined so that the multi-aspect quality vector is a nondecreasing function of its components (quality factors; Kulikowski, 1986). For example, if \( t_0 \) denotes the time of data creation and \( t \) is a current time, then for \( t_0 \leq t \) neither the difference \( t - t_0 \) (as not satisfying the first condition) nor \( t - t_0 \) (as not satisfying the third condition) can be used as data accuracy measures. However, if \( T \), such that \( 0 \leq t_0 \leq T \), denotes the maximum data validity time then for \( t \leq T \) a difference \( \theta = T - t \) or a normalized difference \( \Theta = \theta / T \), both satisfying the above-mentioned conditions, can be used as actuality measures characterizing one of data quality aspects. For similar reasons, if \( \Delta \) is a finite length of admissible data values and \( \delta \) is the length of data confidence interval, where \( 0 < \delta < \Delta \), then \( c = 1 / \delta \) or \( C = \Delta / \delta \), rather than \( \delta \), can be used as data accuracy measures.

Comparison of Multi-Aspect Data Qualities in a Linear Vector Space

For any pair of vectors \( Q', Q'' \) describing multi-aspect qualities of some single data they can be compared according to the semi-ordering rules in the given linear vector space (Akilov & Kukateladze, 1978). Linearity of the vector space admits the operations of vectors adding, multiplying by real numbers and taking the differences of vectors. For comparison of vectors a so-called positive cone \( K^+ \) in the vector space should be established, as shown in Figure 1. It is assumed that \( Q' \) has lower value than \( Q'' \) \((Q' \prec Q'')\) if their difference \( \Delta Q = Q'' - Q' \) is a vector belonging to \( K^+ \). If neither \( Q' \prec Q'' \) nor \( Q'' \prec Q' \) then the vectors \( Q', Q'' \) are called mutually incomparable, meaning that no quality vector in this pair with respect to the other one can be preferred. Figure 1a shows a situation of \( Q' \prec Q'' \), while Figure 1b shows vectors’ incomparability caused by a narrower positive cone (between the dotted lines) being assumed.

The example shows that the criteria of multi-aspect data quality comparability can be changed according to the user’s requirements: They become more rigid when the positive cone \( K^+ \) established is narrower. In particular, if \( K^+ \) becomes as narrow as being close to a positive half-axis, then \( Q' \prec Q'' \) if all components of \( Q' \) are proportional to the corresponding ones of \( Q'' \) with the coefficient of proportionality <1. On the other hand, if \( K^+ \) becomes as large as the positive sector of the vector space, then \( Q' \sim Q'' \) means that all components of \( Q' \) are lower than the corresponding ones of \( Q'' \) in practice the domination of the components of \( Q'' \) over those of \( Q' \) may be not significant, excepting one or several selected components. Between the above-mentioned two extremes there is a large variety of data quality vector semi-ordering, including those used in multi-aspect optimization theory. In particular, if \( Q' = [q_1', q_2', \ldots, q_k'] \) and \( Q'' = [q_1'', q_2'', \ldots, q_k''] \) are two vectors, then it can be defined their scalar product:

\[
(Q', Q'') = q_1' \times q_1'' + q_2' \times q_2'' + \ldots + q_k' \times q_k''
\]

and the length (a norm) of a vector \( Q \) given by the expression:

\[
\| Q \| = \sqrt{(Q, Q)}
\]

Then any fixed positive-components vector \( C \) of a norm \( \| C \| = 1 \) indicates the direction of the \( K^+ \) cone axis. For any given data quality vector \( Q \), an angle \( \angle (C, Q) \) between the vectors \( C \) and \( Q \) can be calculated from its cosine given by:

\[
\cos \angle (C, Q) = (C, Q) / \| Q \|
\]

This expression also can be used for data quality comparison: For the given pair of quality vectors it is assumed that \( Q' \prec Q'' \) if and only if \( \cos \angle (C, Q') < \cos \angle (C, Q'') \); i.e., when \( Q'' \) is closer to \( C \) than \( Q' \). This type of vector comparison leads, in fact, to a comparison of weighted linear combinations of data quality factors.

Remarks on Composite Data Quality Assessment

The above-described single data quality assessment principles should be extended to composite data and higher-order data structures. In general, it is not a trivial problem, the higher-order data structures usually being composed of various types of simple data. The quality of a record

Figure 1. Principles of comparison of vectors describing multi-aspect data values

www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

The Effect of User View Characteristics on Database Design Performance of Novices

www.igi-global.com/article/effect-user-view-characteristics-database/51182?camid=4v1a

Implementation of an Interface to Multiple Databases

www.igi-global.com/article/implementation-interface-multiple-databases/51147?camid=4v1a

SQL Code Poisoning: The Most Prevalent Technique for Attacking Web Powered Databases

www.igi-global.com/chapter/sql-code-poisoning/8025?camid=4v1a

Accelerating Web Service Workflow Execution via Intelligent Allocation of Services to Servers

www.igi-global.com/chapter/accelerating-web-service-workflow-execution/63675?camid=4v1a