Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 1997 Proceedings

Americas Conference on Information Systems (AMCIS)

8-15-1997

Improving Decision Making Using Confidence Scaling for Enhanced Data Quality

Amita Goyal
Virginia Commonwealth University, amita@saturn.vcu.edu

Shirley Becker
American University, sbecker@american.edu

Improving Decision Making Using Confidence Scaling for Enhanced Data Quality

Amita Goyal Chin, Ph.D.
Department of Information Systems
School of Business
Virginia Commonwealth University
Richmond, VA 23284-4000
Email: amita@saturn.vcu.edu
URL: http://pegasus.isy.vcu.edu/~amita
Telephone: (804) 828-7131
Fax: (804) 828-8884
Shirley Becker, Ph.D.
Computer Science & Information Systems
American University
4400 Mass. Ave. N.W.
Washington D.C. 20016
Email: sbecker@american.edu
Telephone: (202) 885-3275
Fax: (301) 349-4633

Abstract

This work identifies measurements of data quality and determines indicators of both data quality and data deficiency. Additionally, this work introduces the concept of confidence scaling which promotes enhanced data quality, resulting in improved decision making.

Introduction

The past decade has witnessed a dramatically increased reliance on data as the central component for informed business decision making processes. The quality of a company's warehoused data therefore directly impacts the quality of the organizational decisions. Unfortunately, many organizations are encumbered with antiquated systems housing erroneous and incomplete data. Johnson et al [Johnson 81] analyzed error characteristics in 55 accounts receivables and 26 inventory audits. Their study found significantly higher error rates in inventory audits, with some systems exhibiting error rates of over 50 percent. In another study surveying 501 medium and large companies, almost two-thirds reported problems resulting from inaccurate, outdated or missing data [Knight 92]. The potential high cost estimates of existing methodologies for identifying and correcting erroneous data dissuade management from pursuing the long term solutions, resulting in either no data quality enhancements or the adoption of some short term and swift bandage solutions. Additional catalysts for insufficient, quick solutions include the perceived embarrassment accompanying the admission of housing error-ridden data and the fear of ensuing lawsuits from customers effected by corporate data inaccuracies. As data continues to manifest intraorganization strategic importance equivalent to that of capital and material, integrating quality standards into data capture processes as well as data maintenance procedures must become a primary focus of strategic organizational efforts.

The purpose of this work is to clearly identify measurements of data quality, determine indicators of data quality and data deficiency, and introduce the concept of confidence scaling for continuous quality screening and maintenance of organizational data. The confidence scaling concept may be used to achieve greater levels of quality in new data systems, as well as to elevate the quality of data in existing information systems.

Data Quality

We consider data quality in electronic data only. Quality of data in non-electronic forms, although necessary, is relatively irrelevant, since most organizational data of any significance is quickly converted to electronic format.

In an absolute sense, whether or not a particular data item is correct is the only characteristic of quality. On a relative scale, however, several factors contribute to the correctness of data. For example, the usefulness of acquired data to the current stream of decision-making affects the quality of the data in a particular task, however, does not affect the data's quality in absolute. To assess the quality of captured data, we consider the following specific characteristics:

Verifiability: In order to obtain any measure of quality at all, the data under consideration must be verifiable as correct. Otherwise, assessment of data quality cannot be made.

Consistency: Consistency of data measures the extent to which records within a database follow the established rules. These include the specific domain restrictions of each field assignment and field tags as well as indexing and editorial decisions. Since most database management systems allow for the integration of data integrity constraints in the specification of relations, appropriate action must be taken to incorporate these constraints into the system and to prohibit the violation of any of these constraints.

Scope/Coverage: The coverage or scope of data determines how well the captured data covers its subject area(s) and if the information is the "authoritative" information in its discipline. Coverage also assesses if there are any serious gaps, including short- or long-term gaps caused by technical problems.

Timeliness: Timeliness must be considered because data has a finite lifecycle. Information considered accurate at a particular point in time may no longer be accurate at a later time. While some captured data may be very resilient to time, most data is highly volatile. Consider, for example, the customer database of a consulting firm. The database schema may include fields for customerName, customerAddress, and customerDateOfBirth. While the accuracy of customerDateOfBirth will be constant over time, customerAddress is likely to change every 2-5 years. The customerName may be resilient for male customers, but will likely change for female consumers. Based on the frequency of updates, timeliness measures the currency of the acquired data.

Error Rate: A high rate of error renders data completely useless for decision making. Even a relatively small rate of error could result in very poor managerial choices, causing organizational losses.

Some indicators of poor quality in data can also be evaluated. These include:

Unsuitability/Irrelevancy: If an organization queries its warehouses and finds a lack of suitable data for decision-making, the organization's data is lacking in quality. Although an organization may store voluminous data, if it is not relevant to the problems at hand, the quality of data is compromised.

Inaccuracy: If incorrect reports and results are outputted from queries of the warehoused data, the merit of the data becomes questionable.

Inconsistency: Inconsistencies in stored data clearly jeopardizes data integrity. Oftentimes, units of components within an umbrella organizations develop or modify and maintain local information systems. While some data remains unique, much of the housed in these systems is duplicated. That is, identical or very similar information is captured in multiple systems. Consider, for example, a university campus consisting of numerous schools and departments. Each departmental unit maintains a local database of student demographic information. The university also maintains this information. A student's request to change personal address information may be submitted only at a unit level. If this information fails to propagate to all other departments storing the student's information, as well as to the university level database, inconsistencies develop in the various information systems.

Confidence Scaling

Organizational dependence on compromised data can result in catastrophic business decision-making. To improve decision-making, data quality must be improved. However, achieving absolute data correctness may not be feasible. Good decision making can exist even with imperfect data; however, this requires a priori knowledge regarding the quality of the data used in the decision making.

To this end, we introduce the notion of confidence scaling. Incorporating a measure of confidence allows decision makers to judge the resulting combined accuracy of the data used in decision making. When data is accumulated, it is collected with a certain level of assurance and reliability. For example, an analysis over time may reveal that an organizational POS scans data with 99. 9% accuracy. Therefore, data collected from an organization's POS system may be assigned the highest level of confidence. Similarly data initially collected from an electronic survey and later verified by clerks may receive a very high scalar for confidence whereas data collected from surveys in a postal mailing and entered into the system by a clerk may be especially prone to errors and hence receive a lower confidence scalar value.

Measuring confidence, by nature, must be specific to an application and to an organization. Just as each organization must develop an information system unique to the business enterprise, each organization must develop and employ a confidence rating scale, such as a Likert scale, specific to their business and their specific policies and procedures. The numeric specifics of the methodology must also be uniquely determined for each organization.

A confidence scaling should be incorporated into each facet of the data lifecycle. It is preferable to capture confidence scaling at the individual data level rather than at a more abstract level, such as a record or table level. This specificity allows the decision-maker more flexibility and access to more detailed confidence information on the warehoused data. Finer granularity also permits for explicit verification.

When any data is retrieved from the system, the associated indices of confidence are also retrieved. A decision-maker can view the additional information of confidence scaling and assign the appropriate weight to the data when making the overall decision. The decision-maker can then decide whether or not to base an important decision on the data. If the data has the highest level of confidence, a decision-maker will be able to confidently use the data in the decision-making process. However, if the decision-maker realizes that the organizational data suffers in confidence, and hence quality, the decision-maker can wisely make the decision at hand without a large regard to the housed data.

In implementing a confidence scalar system, reinspection could occur more methodically than simple random inspection. For example, a module that lowers the confidence rating of particular data elements to reflect the timeliness of the data may be incorporated into the delayed rules based constraints. This means that after a finite period of time following the acquisition of all volatile data elements, the confidence scalar would drop to reflect the timeliness characteristic of data. For example, two years after acquisition of customerAddress information, the confidence scalar would drop. These items could be flagged for reinspection and verified for correctness or updated appropriately.

Conclusion

Organizations are continuously in need of cost-effective methodologies for efficiently harnessing the full potential of their data so as to enhance their competitive business advantage and achieve their strategic corporate goals. The advent and ready availability of a multitude of data warehouses and other various data sources have further deemed it imperative for organizations to develop and adhere to methodologies for effectively managing the quality of vasts volumes of information diverse in form as well as in origin. The ability and degree to which the organizations can accomplish strategic goals, however, is a direct outcome of the quality of data and the implemented procedures and policies for quality assurance within the organization. Regardless of the multitudes of filtration policies and procedures in place, every organization must implement and exercise procedures for search and correction of errors. The absence of such efforts

compromises data quality. In this work, we have introduced the concept of confidence scaling. Attaching such a scalar to all acquired data allows the decision maker to better understand the quality of the data being used and to place the appropriate importance on the data when making a decision.

References

Johnson, Johnny R. , Leitch, R. A. , and Neter, J. "Characteristics of errors in accounts receivable and inventory audits." The Accounting Review. Vol. LVI, No. 2, April 1981.

Knight, Bob. "The data pollution problem." Computerworld. September 28, 1992.