Data Quality in Cooperative Information Systems

Carlo Marchetti
Università di Roma "La Sapienza", Italy

Massimo Mecella
Università di Roma "La Sapienza", Italy

Monica Scannapieco
Università di Roma "La Sapienza", Italy

Antonino Virgillito
Università di Roma "La Sapienza", Italy

INTRODUCTION

A Cooperative Information System (CIS) is a large-scale information system that interconnects various systems of different and autonomous organizations, geographically distributed and sharing common objectives (De Michelis et al., 1997). Among the different resources that are shared by organizations, data are fundamental; in real world scenarios, organization A may not request data from organization B, if it does not trust B's data (i.e., if A does not know that the quality of the data that B can provide is high). As an example, in an e-government scenario in which public administrations cooperate in order to fulfill service requests from citizens and enterprises (Batini & Mecella, 2001), administrations very often prefer asking citizens for data rather than from other administrations that have stored the same data, because the quality of such data is not known. Therefore, lack of cooperation may occur due to lack of quality certification.

Uncertified quality also can cause a deterioration of the data quality inside single organizations. If organizations exchange data without knowing their actual quality, it may happen that data of low quality will spread all over the CIS. On the other hand, CISs are characterized by high data replication (i.e., different copies of the same data are stored by different organizations). From a data quality perspective, this is a great opportunity; improvement actions can be carried out on the basis of comparisons among different copies in order to select the most appropriate one or to reconcile available copies, thus producing a new improved copy to be notified to all interested organizations.

In this article, we describe possible solutions to data quality problems in CISs that have been implemented within the DaQuinCIS architecture. The description of the architecture will allow us to understand the challenges posed by data quality management in CISs. Moreover, the presentation of the DaQuinCIS services will provide examples of techniques that can be used to face data quality challenges and will show future research directions that should be investigated.

BACKGROUND

Data quality traditionally has been investigated in the context of single information systems. Only recently, there is a growing attention toward data quality issues in the context of multiple and heterogeneous information systems (Berti-Equille, 2003; Bertolazzi & Scannapieco, 2001; Naumann et al., 1999). In cooperative scenarios, the main data quality issues regard (Bertolazzi & Scannapieco, 2001):

- assessment of the quality of the data owned by each organization; and
- methods and techniques for exchanging and improving quality information.

For the assessment issue, some of the results already achieved for traditional systems can be borrowed. In the statistical area, a lot of work has been done since the late 1960s. Record linkage techniques have been proposed, most of them based on the Fellegi and Sunter (1969) model. Also, edit imputation methods based on the Fellegi and Holt (1976) model have been provided in the same area. Instead, in the database area, record matching techniques (Hernandez & Stolfo, 1998) and data cleaning tools (Galhardas et al., 2000) have been proposed as a contribution to data quality assessment. In Winkler (2004), a survey of data quality assessment

techniques is provided, covering both statistical techniques and data cleaning solutions.

When considering the issue of exchanging data and the associated quality, a model to export both data and quality data needs to be defined. Some conceptual models to associate quality information to data have been proposed, which include an extension of the entity-relationship model (Wang et al., 1993) and a data warehouse conceptual model with quality features described through the description logic formalism (Jarke at al., 1995). Both models are for a specific purpose: the former to introduce quality elements in relational database design; the latter to introduce quality elements in the data warehouse design. In Mihaila et al. (2000), the problem of the quality of Web-available information has been faced in order to select data with high quality coming from distinct sources; every source has to evaluate some pre-defined data quality parameters and to make their values available through the exposition of metadata.

Furthermore, exchanging data in CISs poses important problems that have been addressed by the data integration literature. Data integration is the problem of combining data residing at different sources and providing the user with a unified view of these data (Lenzerini, 2002). As described in Yan et al. (1999), when performing data integration, two different types of conflicts may arise: semantic conflicts, due to heterogeneous source models; and instance-level conflicts, due to what happens when sources record inconsistent values on the same objects. The data quality broker described in the following is a system solving instance-level conflicts. Other notable examples of data integration systems within the same category are AURORA (Yan et al., 1999) and the system described in Sattler et al. (2003). AURORA supports conflict tolerant queries (i.e., it provides a dynamic mechanism to resolve conflicts by means of defined conflict resolution functions). The system described in Sattler et al. (2003) describes how to solve both semantic and instance-level conflicts. The proposed solution is based on a multi-database query language, called FraQL, which is an extension of SQL with conflict resolution mechanisms. A system that also takes into account metadata for instance-level conflict resolution is described in Fan et al. (2001). Such a system adopts the ideas of the context interchange framework (Bressan et al., 1997); therefore, context-dependent and independent conflicts are distinguished, and, accordingly, to this very specific direction, conversion rules are discovered on pairs of systems.

Finally, among the techniques explicitly proposed to perform query answering in CISs, we cite Naumann et al. (1999), in which an algorithm to perform query planning based on the evaluation of data sources' qualities, specific queries' qualities, and query results' qualities is described.

MAIN THRUST

In current government and business scenarios, organizations start cooperating in order to offer services to their customers and partners. Organizations that cooperate have business links (i.e., relationships, exchanged documents, resources, knowledge, etc.) connecting each other. Specifically, organizations exploit business services (e.g., they exchange data or require services to be carried out) on the basis of business links, and, therefore, the network of organizations and business links constitutes a cooperative business system. As an example, a supply chain, in which some enterprises offer basic products and some others assemble them in order to deliver final products to customers, is a cooperative business system. As another example, a set of public administrations, which need to exchange information about citizens and their health state in order to provide social aids, is a cooperative business system derived from the Italian e-government scenario (Batini & Mecella, 2001).

A cooperative business system exists independently of the presence of a software infrastructure supporting electronic data exchange and service provisioning. Indeed, cooperative information systems are software systems supporting cooperative business systems; in the remaining of this article, the following definition of CIS is considered:

A cooperative information system is formed by a set of organizations that cooperate through a communication infrastructure N, which provides software services to organizations as well as reliable connectivity. Each organization is connected to N through a gateway G, on which software services offered by the organization to other organizations are deployed. A user is a software or human entity residing within an organization and using the cooperative system.

Several CISs are characterized by a high degree of data replicated in different organizations; for example, in an e-government scenario, the personal data of a citizen are stored by almost all administrations. But in such scenarios, the different organizations can provide the same data with different quality levels; thus, any user of data may appreciate to exploit the data with the highest quality level, among the provided ones.

Therefore, only the highest quality data should be returned to the user, limiting the dissemination of low quality data. Moreover, the comparison of the gathered data values might be used to enforce a general improvement of data quality in all organizations.

In the context of the DaQuinCIS project[1], we are proposing an architecture for the management of data