Data Quality Improvement in the DaQuinCIS System

Luca De Santis, Diego Milano, Gabriele Palmieri
Monica Scannapieco and Tiziana Catarci

Technical Report n. 14
2004

Dipartimento di Informatica e Sistemistica “Antonio Ruberti”
Università degli Studi di Roma “La Sapienza”
Via Eudossiana, 18 - 00184 Roma
Via Buonarroti, 12 - 00185 Roma
Via Salaria, 113 - 00198 Roma
www.dis.uniroma1.it
Data Quality Improvement in the DaQuinCIS System

Luca De Santis, Diego Milano, Gabriele Palmieri, Monica Scannapieco and Tiziana Catarci

abstract

Data quality improvement is becoming an increasingly important issue. In contexts where data are replicated among different sources, data quality improvement is possible through extensive data comparisons: whereas copies of same data are different because of data errors, comparisons help to reconcile such copies. Best quality copies can be selected or constructed in order to correct other copies. Record matching algorithms can support the task of linking different copies of the same data in order to engage reconciliation activities; for instance, a periodical running of record matching algorithms can be performed in order to reconcile copies with different quality. Nevertheless, the extensive running of such algorithms is typically performed in fixed instants. This allows for periods in which the quality of data can deteriorate, while no quality improvement action is performed on data. In this paper, we describe the DaQuinCIS platform for data quality improvement in contexts where data are replicated among heterogeneous and distributed sources. The quality improvement strategy underlying the proposed platform complements a periodical record matching activity with an “on-line” quality improvement, performed at query processing time. We experimentally show the feasibility and effectiveness of our approach by applying it to real databases; we also quantitatively evaluate the efficiency of our system.

1 Introduction

Data quality is more than simply data accuracy. It is a complex concept defined by other dimensions such as currency, completeness, consistency; such dimensions often depend from the context where data are used and also from specific users within a given context [24]. It follows that it’s not easy to obtain good quality data. Despite this complexity, the crucial role of data requires us to face and solve data quality problems. In specific environments
characterized by extensive data replication, high quality of data becomes a strict requirement: poor quality data means that conflicts arise thus making data quality problems even more urgent to be solved.

1.1 Data Quality & CISs

We define Cooperative Information Systems (CISs) as all distributed and heterogeneous information systems that cooperate by sharing information, constraints, and goals [17]. CISs include data integration systems as well as systems that share common information while not explicitly integrating data. We argue that quality of data is a necessary requirement for a CIS, i.e. CISs need data quality. First, a system in the CIS will not easily exchange data with another system without a knowledge on their quality, and cooperation becomes difficult without data exchanges. Second, when poor quality data are exchanged, there is a progressive deterioration of data stored in the whole CIS. Indeed, uncontrolled exchanges of low quality data cause a diffusion of such data throughout the system, thus lowering the quality of each data asset in the CIS. Third, when a CIS is a data integration system, data integration itself cannot be performed if data quality problems are not fixed. As an example, results of queries executed over local sources must be reconciled and merged, and quality problems resulting from a comparison of results need to be solved in order to provide the data integration system with the required information [5].

On the other hand, data replication in the CIS can be exploited for improving data quality, i.e. data quality needs CISs. Though it is possible to enact quality improvement actions by simply looking at single sources, there are some limitations that cannot be overcome by having only one copy of data to look at. For instance, accuracy improvement can be performed by having syntactic dictionaries as references; therefore, values for a field name could be quite easily syntactically checked. Nevertheless, things can become more difficult if field values cannot be compared with reference dictionaries; for instance, this is often the case of numerical data. Furthermore, even in the case in which syntactical checks can be applied, the semantic problem is not solved; for instance,
if a date is syntactically correct, i.e. it respects a given format, but it is actually the wrong date, this is not easily detected by simply looking at one source. Instead, in CISs different copies of the same data are typically stored by multiple sources and can be compared in order to detect quality problems and possibly solve them.

1.2 Data Quality Improvement in DaQuinCIS

In this paper, we propose an overall approach to data quality improvement in cooperative information systems, called the DaQuinCIS (Data Quality in Cooperative Information Systems) approach \(^1\). The approach is supported by a platform implemented as a set of peer to peer services for quality improvement and maintenance in CISs. The basic idea of the DaQuinCIS approach is to improve quality of data replicated in a CIS through extensive data comparisons: whereas copies of same data are different because of data errors, comparisons help to reconcile such copies.

The DaQuinCIS approach proposes to improve data quality in two distinct steps, namely a periodical record matching and a query-time quality improvement. The approach has been experimented on real data bases owned by the Italian Public Administration.

1.3 Outline of the Paper

The paper is organized as follows. Section 2 provides an overview of the DaQuinCIS architecture and of the DaQuinCIS approach to quality improvement. Section 3 describes the Record Matcher component of the architecture and the algorithm that such module implements. Section 4 describes the Query Processor component of the architecture and the new approach to quality improvement at query-time used in DaQuinCIS. Section 5 discusses the implemented system. Section 6 describes our experimental results.

\(^1\)The DaQuinCIS approach has been developed in the context of the project “DaQuinCIS - Methodologies and Tools for Data Quality inside Cooperative Information Systems” (http://www.dis.uniroma1.it/~dq/).
Section 7 discusses related work, and Section 8 describes some additional issues and concluding remarks.

2 The DaQuinCIS Architecture

The DaQuinCIS architecture allows to exchange data and associated quality and exploits data replication to improve the overall quality of cooperative data.

Each organization offers services to other organizations and also specific services to its internal back-end systems. Such services are deployed on cooperative gateways that interface each cooperating organization with the other ones. Moreover, the communication infrastructure itself offers some specific services. Services deployed on cooperative gateways are all identical and peer, i.e., they are instances of the same software artifacts, and act both as servers and clients of the other peers depending on the specific activities to be carried out. The overall architecture is depicted in Figure 1.

Organizations export data and quality data according to a common model, referred to as Data and Data Quality ($D^2Q$) model. It includes the definitions of (i) constructs to represent data, (ii) a common set of data quality properties, (iii) constructs to represent them and (iv) the association between data and quality data. More details on the $D^2Q$ model can be found in [15].

In order to produce data and quality data according to the $D^2Q$ model, each organization deploys on its cooperative gateway a Quality Factory service that is responsible for evaluating the
quality of its own data. More details on the Quality Factory can be found in [7].

The **Data Quality Broker** poses, on behalf of a requesting user, a data request over other cooperating organizations, also specifying a set of quality requirements that the desired data have to satisfy; this is referred to as *quality brokering functionality*. Details on the brokering functionality can be found in [15].

In this paper, we focus on a different functionality of the Data Quality Broker, namely the *quality improvement functionality* that is carried on in two phases: *(i)* record matching is periodically performed by the *Record Matcher* component; *(ii)* comparisons among same data are performed at query time by the *Query Processor*. Different copies of the same data received as responses to a query are reconciled and a best-quality value is selected and proposed to organizations.

The **Quality Notification Service** is a publish/subscribe engine used as a quality message bus between services and/or organizations. More specifically, it allows quality-based subscriptions for users to be notified on changes of the quality of data. For example, an organization may want to be notified if the quality of some data it uses degrades below a certain threshold, or when high quality data are available. Also the Quality Notification Service is deployed as a peer-to-peer system. More details on the Quality Notification Service can be found in [21].

The **Rating Service** associates trust values to each data source in the CIS. These values are used to determine how much an organization can be trusted with respect to provided data. The interested reader can refer to [8] for more details.

### 2.1 The DaQuinCIS Approach to Data Quality Improvement in CISs

The DaQuinCIS approach proposes to improve data quality in two distinct steps:

- **Periodical record matching**: a record matching algorithm is run on data sets stored by the organizations in the CIS. The analyzed data sets are the ones that cooperating organiza-
tions exchange with each other within cooperative processes to which they participate [2].

- **Query-time improvement**: data comparisons are performed at query processing time by comparing data received as answers to queries.

The DaQuinCIS approach allows to improve data quality in a continuous fashion. The application of record matching algorithms is indeed limited to fixed instants. This allows for periods in which the quality of data can deteriorate, while no quality improvement action is performed on data. The DaQuinCIS approach complements a periodical record matching activity with an “online” quality improvement, performed at query processing time.

The periodical record matching is performed by an algorithm that is implemented by the Record Matcher. Such algorithm was first proposed in [3] and is summarized in Section 3.

The query-time quality improvement is first introduced in this paper and adds a semantics to perform quality improvement to the Data Quality Broker. The Data Quality Broker is in essence a peer-to-peer data integration system. The quality improvement functionality is embedded in the query processing and implies the definition of a mapping of the global schema concepts in terms of concepts at local sources in order to retrieve all copies of the same data in the system, compare and improve them. The query-time quality improvement uses the record matching algorithm in order to perform comparisons on results of queries. Section 4 describes the newly introduced quality improvement functionality as performed by the Query Processor.

### 3 The Record Matcher

Record Matching, also known as Record Linkage [9] or Object Identity problem [23], is the problem of identifying if two records are related to the same real world entity. In the DaQuinCIS platform, a record matching activity is performed in two phases:

- a periodical record matching is run in order to align different copies of the same entities that are present in data sources.