Data Quality Evaluation in Data Integration Systems

The needs of accessing in a uniform way to information available in multiple data sources are increasingly higher and generalized, particularly in the context of decision making applications which need a comprehensive analysis and exploration of data. With the development of Data Integration Systems (DIS), information quality is becoming a *first class* property which is more and more required by end-users.

This thesis deals with data quality evaluation in DIS. Specifically, we address the problems of evaluating the quality of the data conveyed to users in response to their queries and verifying if users' quality expectations can be achieved. We also analyze how quality measures can be used for improving the DIS and enforcing data quality. Our approach consists in studying one quality factor at a time, analyzing its impact within a DIS, proposing techniques for its evaluation and proposing improvement actions for its enforcement. Among the quality factors that have been proposed, this thesis analyzes two main ones: *data freshness* and *data accuracy*.

We analyze the different definitions and metrics proposed for data freshness and data accuracy and we abstract the properties of the DIS that impact on their evaluation. We summarize the analysis of each factor with a taxonomy, which allows comparing existent works and highlighting open problems.

We propose a quality evaluation framework that models the different elements involved in data quality evaluation, namely: data sources, user queries, DIS processes, DIS properties, quality measures and quality evaluation algorithms. In particular, DIS processes are modeled as workflow processes in which the workflow activities perform the different tasks that extract, integrate and convey data to end-users. Our reasoning support for quality evaluation is a direct acyclic graph, called quality graph, which has the same workflow structure than the DIS and contains, as labels, the DIS properties that are relevant for quality evaluation. We develop quality evaluation algorithms that take as input source data quality values and DIS property values and combine such values obtaining a value for the data conveyed by the DIS. They are based on the graph representation and combine property values while traversing the graph. Evaluation algorithms can be instantiated for taking into account the properties that influence data quality in a particular application. The idea behind the framework is to define a flexible context which allows specializing evaluation algorithms for specific application scenarios.

The quality values obtained during data quality evaluation are compared to those expected by users. If quality expectations are not satisfied, several improvement actions can be taken. We suggest some elementary improvement actions that can be composed to improve data quality in concrete DISs. For enforcing data freshness, we propose the analysis of the DIS at different abstraction levels in order to identify its critical points (the portions of the DIS that cause the non-achievement of freshness expectations) and to target the study of improvement actions for these points. For enforcing data accuracy, we propose the partitioning of query result in areas (some attributes, some tuples) having homogeneous accuracy. This allows user applications to retrieve only the most accurate data, to filter data not satisfying an accuracy threshold or to incrementally convey areas (e.g. displaying first the most accurate areas). Our approach differentiates from existing source selection proposals because we allow the selection of the areas having the best accuracy instead of only selecting whole relations.

The main contributions of this thesis are: (i) a detailed analysis of data freshness and data accuracy quality factors; (ii) the proposal of techniques and algorithms for the evaluation and enforcement of data freshness and data accuracy; and (iii) a prototype of a quality evaluation tool oriented to be used in practical contexts of DIS management.