

Université de Versailles Saint-Quentin en Yvelines
Versailles, FRANCE

Universidad de la República
Montevideo, URUGUAY

PhD THESIS

Presented by
Verónica PERALTA

For obtaining the degree of PhD Doctor
in Informatics

Thesis Subject:

Data Quality Evaluation in Data Integration Systems

Abstract

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 of the most used 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. Among these elements there are data sources, user queries, DIS processes, DIS properties, quality measures and quality evaluation algorithms. In particular, we model DIS processes as workflow processes in which the workflow activities perform the different tasks that extract, integrate and convey data to end-users. 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 workflow graph representation and combine property values while traversing the graph. The idea behind the framework is to define a flexible context which allows specializing evaluation algorithms in order to take into account the properties of 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.

For enforcing data freshness, we propose an enforcement approach that supports the analysis of the DIS at different abstraction levels in order to identify its weak points (the portions of the DIS that are the bottlenecks for achieving freshness expectations) and to target the study of improvement actions for these weak points. The graph representation of the DIS allows the rapid visualization of such weak points. The proposed improvement actions are building blocks that can be composed to improve data freshness in concrete DISs.

For enforcing data accuracy, we propose the partitioning of query result in areas (some attributes of 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 and if user wants to see more, i.e. the result is not complete enough, displaying him the following areas). This represents an improvement to existing source selection proposals because accurate areas of several source relations can be combined while discarding inaccurate areas (instead of discarding whole sources).

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.

Résumé

Les besoins d'accéder, de façon uniforme, à des sources de données multiples, sont chaque jour plus forts, particulièrement, dans les systèmes décisionnels qui ont besoin d'une analyse compréhensive des données. Avec le développement des Systèmes d'Intégration de Données (SID), la qualité de l'information est devenue une propriété de premier niveau de plus en plus exigée par les utilisateurs.

Cette thèse porte sur la qualité des données dans les SID. Nous nous intéressons, plus précisément, aux problèmes de l'évaluation de la qualité des données délivrées aux utilisateurs en réponse à leurs requêtes et de la satisfaction des exigences des utilisateurs en terme de qualité. Nous analysons également l'utilisation de mesures de qualité pour l'amélioration de la conception du SID et de la qualité des données. Notre approche consiste à étudier un facteur de qualité à la fois, en analysant sa relation avec le SID, en proposant des techniques pour son évaluation et en proposant des actions pour son amélioration. Parmi les facteurs de qualité qui ont été proposés, cette thèse analyse deux facteurs de qualité : *la fraîcheur* et *l'exactitude* des données.

Nous analysons les différentes définitions et mesures qui ont été proposées pour la fraîcheur et l'exactitude des données et nous faisons émerger les propriétés du SID qui ont un impact important sur leur évaluation. Nous résumons l'analyse de chaque facteur par le biais d'une taxonomie, qui sert à comparer les travaux existants et à faire ressortir les problèmes ouverts.

Nous proposons un canevas qui modélise les différents éléments liés à l'évaluation de la qualité tels que les sources de données, les requêtes utilisateur, les processus d'intégration du SID, les propriétés du SID, les mesures de qualité et les algorithmes d'évaluation de la qualité. En particulier, nous modélisons les processus d'intégration du SID comme des processus de workflow, dans lequel les activités réalisent les tâches qui extraient, intègrent et envoient des données aux utilisateurs. Nous développons des algorithmes d'évaluation qui prennent en entrée les valeurs de qualité des données sources et les propriétés du SID, et, combinent ces valeurs pour qualifier les données délivrées par le SID. Ils se basent sur la représentation en forme de graphe du workflow et combinent les valeurs des propriétés en traversant le graphe. L'idée derrière le canevas est de définir un contexte flexible qui permet la spécialisation des algorithmes d'évaluation pour tenir compte des propriétés de scénarios d'application.

Les valeurs de qualité obtenues pendant l'évaluation sont comparées à celles attendues par les utilisateurs. Des actions d'amélioration peuvent se réaliser si les exigences de qualité ne sont pas satisfaites.

Notre approche pour améliorer la fraîcheur des données consiste à l'analyse du SID à différents niveaux d'abstraction, de façon à identifier ses points faibles et cibler l'application d'actions d'amélioration de ces points-là. Les actions d'amélioration proposées sont des briques de base qui peuvent être composées pour améliorer la fraîcheur dans un SID concret.

Notre approche pour améliorer l'exactitude des données consiste à partitionner les résultats des requêtes en portions ayant une exactitude homogène. Cela permet aux applications utilisateur de visualiser seulement les données les plus exactes, de filtrer les données ne satisfaisant pas les exigences d'exactitude ou de visualiser les données par tranche selon leur exactitude. Comparée aux approches existantes de sélection de sources, notre proposition permet de filtrer les portions les plus imprécises au lieu de filtrer des sources entières.

Les contributions principales de cette thèse sont : (1) une analyse détaillée des facteurs de qualité fraîcheur et exactitude ; (2) la proposition de techniques et algorithmes pour l'évaluation et l'amélioration de la fraîcheur et l'exactitude des données ; et (3) un prototype d'évaluation de la qualité utilisable dans la conception de SID.

Content

CHAPTER 1. INTRODUCTION	1
1. CONTEXT	1
2. MOTIVATIONS AND PROBLEMS.....	2
3. OUR PROPOSITION	4
3.1. <i>Technical issues addressed in this thesis</i>	4
3.2. <i>Main contributions</i>	5
4. OUTLINE OF THE THESIS	5
CHAPTER 2. STATE OF THE ART	7
1. INTRODUCTION	7
2. DATA FRESHNESS	7
2.1. <i>Freshness definitions</i>	8
2.2. <i>Freshness measurement</i>	8
2.3. <i>Dimensions for freshness analysis</i>	10
2.4. <i>A taxonomy for freshness measurement techniques</i>	12
2.5. <i>Some systems that consider data freshness</i>	13
2.6. <i>Research problems</i>	15
3. DATA ACCURACY	18
3.1. <i>Accuracy definitions</i>	19
3.2. <i>Accuracy measurement</i>	22
3.3. <i>Dimensions for accuracy analysis</i>	25
3.4. <i>A taxonomy for accuracy measurement techniques</i>	31
3.5. <i>Some systems that consider data accuracy</i>	33
3.6. <i>Research problems</i>	35
4. CONCLUSION.....	37
CHAPTER 3. DATA FRESHNESS.....	39
1. INTRODUCTION	39
2. DATA QUALITY EVALUATION FRAMEWORK.....	41
2.1. <i>Definition of the framework</i>	41
2.2. <i>The approach for data quality evaluation in data integration systems</i>	44
3. DATA FRESHNESS EVALUATION	45
3.1. <i>Basic evaluation algorithm</i>	46
3.2. <i>Overview of the instantiation approach</i>	48
3.3. <i>Modeling of scenarios</i>	50
3.4. <i>Identification of appropriate properties</i>	51
3.5. <i>Instantiation of the evaluation algorithm</i>	54
3.6. <i>Propagation of freshness expectations</i>	55
3.7. <i>Usages of the approach</i>	57
4. DATA FRESHNESS ENFORCEMENT.....	60
4.1. <i>Top-down analysis of data freshness</i>	60
4.2. <i>Browsing among quality graphs</i>	67
4.3. <i>Determination of critical paths</i>	69
4.4. <i>Improvement actions</i>	73
4.5. <i>Summarizing example</i>	79
5. SYNCHRONIZATION OF ACTIVITIES.....	81
5.1. <i>DIS synchronization problem</i>	82
5.2. <i>Characterization of the solution space</i>	83
5.3. <i>Solutions to the DIS synchronization problem</i>	86
6. CONCLUSION.....	90

CHAPTER 4. DATA ACCURACY	91
1. INTRODUCTION	91
2. INTUITIVE APPROACH.....	93
3. BACKGROUND.....	96
3.1. <i>Some related approaches for accuracy evaluation</i>	96
3.2. <i>Query rewriting</i>	99
3.3. <i>Selectivity estimation</i>	100
3.4. <i>Quality evaluation framework</i>	101
4. FORMAL APPROACH	102
4.1. <i>Partitioning of source relations according to accuracy homogeneity</i>	103
4.2. <i>Rewriting of user queries in terms of partitions</i>	106
4.3. <i>Estimation of data accuracy of query results</i>	108
4.4. <i>Reuse of the quality evaluation framework</i>	111
5. ACCURACY IMPROVEMENT	114
6. CONCLUSION.....	116
CHAPTER 5. EXPERIMENTATION AND APPLICATIONS.....	117
1. INTRODUCTION	117
2. PROTOTYPE	117
2.1. <i>Functionalities</i>	118
2.2. <i>Architecture</i>	119
2.3. <i>Interface</i>	120
2.4. <i>Practical use of the tool</i>	121
2.5. <i>Liberation of versions</i>	122
3. APPLICATIONS.....	122
3.1. <i>An adaptive system for aiding in the generation of mediation queries</i>	122
3.2. <i>Evaluating data freshness in a web warehousing application</i>	127
3.3. <i>Evaluating data accuracy in a data warehousing application</i>	133
4. EVALUATION OF PERFORMANCE AND LIMITATIONS OF THE DQE TOOL	135
4.1. <i>Generation of test data sets</i>	136
4.2. <i>Test of limitations</i>	140
4.3. <i>Test of performance</i>	141
5. CONCLUSION.....	144
CHAPTER 6. CONCLUSIONS AND PERSPECTIVES	145
1. SUMMARY AND CONTRIBUTIONS.....	145
2. PERSPECTIVES.....	146
2.1. <i>Near future work</i>	147
2.2. <i>Other research perspectives</i>	148
2.3. <i>Towards quality-driven design of DIS</i>	150
ANNEX A. DESIGN OF THE DQE TOOL	153
1. DATA MODEL	153
2. METABASE.....	155
ANNEX B. INSTANTIATION OF THE FRESHNESS EVALUATION ALGORITHM.....	157
1. MEDIATION APPLICATION SCENARIO.....	157
2. WEB WAREHOUSING APPLICATION SCENARIO	159
REFERENCES.....	163