Quality Awareness in Data Management and Mining

Laure BERTI-ÉQUILLE

Habilitation à Diriger des Recherches
IRISA - Université de Rennes 1

25 Juin 2007

IRISA

UNIVERSITÉ DE RENNES 1

Outline

1. Activities
2. Problem statement
3. Metadata management
4. Data mining
5. Applications
6. Conclusions

Plan

1. **Activities**
   - Education and qualification
   - Teaching activities
   - Research activities
   - Projects, contracts and collaborations
   - Organization activities

Doctoral Qualification

1996: Université de Paris IX-Dauphine

- Master's Degree in Computer Science

1996-1999: Université de Toulon et du Var

- Ph.D. in Computer Science : "*Qualité des données et leur recommandation: application à la veille technologique*"
- "*Moniteur C.I.E.S.*"

Post-doctoral Position

1999-2000: Université d'Avignon et Pays du Vaucluse

- Assistant Professor

Current Position

2000 - now: Université de Rennes 1 - IRISA

- Associate Professor

Teaching activities

Courses at Université de Rennes 1

Databases	DIIC2 & MPRO2
Advanced Databases	MPRO2 TC
Data Warehouses	MPRO2 MIAGE
XML Technologies	MPRO1 MIAGE
Object-Oriented System Design	MPRO2 MIAGE
Project Management	MPRO1-2 MIAGE

Details available at http://www.irisa.fr/Laure.Berti-Equille/Enseignement.html

Numbers

Publications since 1996

2 book chapters and 3 edited proceedings
5 papers in intl. journals et 7 in national journals
15 papers in intl. conferences and 6 in intl. workshops
7 papers in national conferences et 2 in national workshops

53% as a unique author

Supervision

1 Graduated Ph.D. and one current Ph.D. student
1 Expert engineer
1 Current post-doc
4 M.S. students and one internship
2 Participations as a reviewer in a Ph.D. jury

Numbers

Publications since 1996

2 book chapters and 3 edited proceedings
5 papers in intl. journals et 7 in national journals
15 papers in intl. conferences and 6 in intl. workshops
7 papers in national conferences et 2 in national workshops

53% as a unique author

Supervision

1 Graduated Ph.D. and one current Ph.D. student
1 Expert engineer
1 Current post-doc
4 M.S. students and one internship
2 Participations as a reviewer in a Ph.D. jury

Coordination

- **European Integrated Project (PF-6)**
  ENTHRONE Phase 1, 2003-2005, Coordinator for INRIA Rennes
- **International Projects**
  - CLINIQ, PHC Italy, Università La Sapienza - IStat, 2006
  - M4, PHC Japan, National Institute of Informatics, 2002
- **National Project (ANR)**
  QUADRIS, ANR-05-MMSA, Coordinator, 2006-2009

Contracts and Collaborations

- **Scientific Responsability**
  - Contract with Genielog, 2005-2006
  - Contract with Écoles Militaires de Coëtquidan, 2003-2008
- **Participation**
  Inter-EPST Project with INSERM U522, 2002-2003

Coordination

- **European Integrated Project (PF-6)**
  ENTHRONE Phase 1, 2003-2005, Coordinator for INRIA Rennes
- **International Projects**
  - CLINIQ, PHC Italy, Università La Sapienza - IStat, 2006
  - M4, PHC Japan, National Institute of Informatics, 2002
- **National Project (ANR)**
  QUADRIS, ANR-05-MMSA, Coordinator, 2006-2009

Contracts and Collaborations

- **Scientific Responsability**
  - Contract with Genielog, 2005-2006
  - Contract with Écoles Militaires de Coëtquidan, 2003-2008
- **Participation**
  Inter-EPST Project with INSERM U522, 2002-2003

Organization

- **Two first editions of the national workshop**
  Data and Knowledge Quality (DKQ)
  in conjunction with EGC, Paris and Lille, January 2005 and 2006
- **Second edition of the international workshop**
  Information Quality in Information Systems (IQIS)
  in conjunction with ACM SIGMOD, Baltimore, USA, June 2005

Participation

- Organization Committee Member:
  BDA'05, JOBIM'02, EDD'01, INFORSID'98
- Program Committee Member:
  21 participations since 2005 including VLDB'07
- Editorial Board Member of two international journals:
  - *International Journal of Information Quality (IJIQ)*
  - *Journal of Digital Information Management (JDIM)*

Organization

- **Two first editions of the national workshop**
  Data and Knowledge Quality (DKQ)
  in conjunction with EGC, Paris and Lille, January 2005 and 2006
- **Second edition of the international workshop**
  Information Quality in Information Systems (IQIS)
  in conjunction with ACM SIGMOD, Baltimore, USA, June 2005

Participation

- Organization Committee Member:
  BDA'05, JOBIM'02, EDD'01, INFORSID'98
- Program Committee Member:
  21 participations since 2005 including VLDB'07
- Editorial Board Member of two international journals:
  - *International Journal of Information Quality (IJIQ)*
  - *Journal of Digital Information Management (JDIM)*

Plan

2 **Problem statement**
- General remarks
- Context of research
- Research axis

Main Data Quality Problems

At the schema level

- X Missing values
- X Domain constraints violation
- X Referential integrity constraints violation
- X Exact duplicates

- X Erroneous categorical data
- X Out-of-date data
- X Inconsistencies
- X Naming conflicts
- X Structural conflicts

Main Data Quality Problems

At the schema level

- √ Missing values
- √ Domain constraints violation
- √ Referential integrity constraints violation
- √ Exact duplicates

- X Erroneous categorical data
- X Out-of-date data
- X Inconsistencies
- X Naming conflicts
- X Structural conflicts

Main Data Quality Problems

At the schema level

√ Missing values

√ Domain constraints violation

√ Referential integrity constraints violation

√ Exact duplicates

X Erroneous categorical data

X Out-of-date data

X Inconsistencies

X Naming conflicts

X Structural conflicts

Main Data Quality Problems

At the instance level

- X Non standardized data
- X Incomplete data
- X Erroneous data and outliers
- X Typos
- X Embedded values
- X Misfielded values
- X Ambiguous or contradictory data
- X Approximate duplicates

Main Data Quality Problems

At the instance level

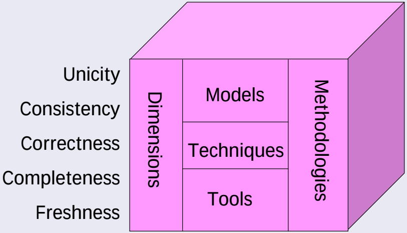
- X Non standardized data
- X Incomplete data
- X Erroneous data and outliers
- X Typos
- X Embedded values
- X Misfielded values
- X Ambiguous or contradictory data
- X Approximate duplicates

Data Quality Research

Convergence of Several Fields

- Statistics
- Databases and Information Systems
- Project and workflow management
- Knowledge engineering

With 5 modalities



Unicity

Consistency

Correctness

Completeness

Freshness

Dimensions

Models

Techniques

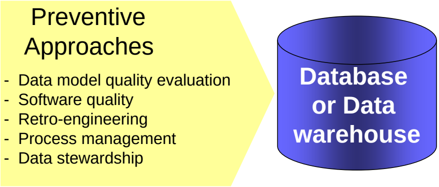
Tools

Methodologies

Main Approaches

Database
or Data
warehouse

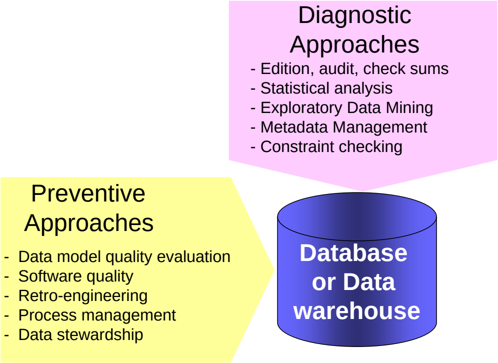
Main Approaches



Preventive Approaches

- Data model quality evaluation
- Software quality
- Retro-engineering
- Process management
- Data stewardship

Main Approaches

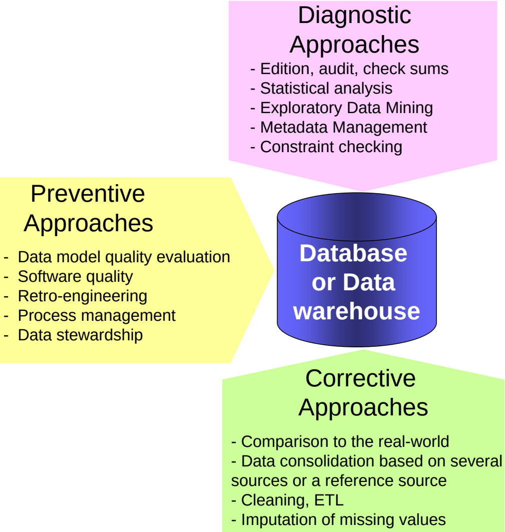


Diagnostic Approaches
- Edition, audit, check sums
- Statistical analysis
- Exploratory Data Mining
- Metadata Management
- Constraint checking

Preventive Approaches
- Data model quality evaluation
- Software quality
- Retro-engineering
- Process management
- Data stewardship

Database or Data warehouse

Main Approaches



Diagnostic Approaches

- Edition, audit, check sums
- Statistical analysis
- Exploratory Data Mining
- Metadata Management
- Constraint checking

Preventive Approaches

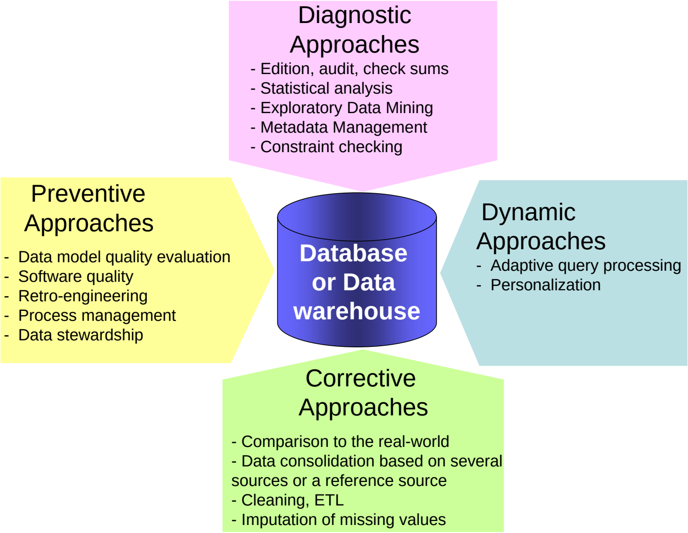
- Data model quality evaluation
- Software quality
- Retro-engineering
- Process management
- Data stewardship

Database or Data warehouse

Corrective Approaches

- Comparison to the real-world
- Data consolidation based on several sources or a reference source
- Cleaning, ETL
- Imputation of missing values

Main Approaches



Diagnostic Approaches

- Edition, audit, check sums
- Statistical analysis
- Exploratory Data Mining
- Metadata Management
- Constraint checking

Preventive Approaches

- Data model quality evaluation
- Software quality
- Retro-engineering
- Process management
- Data stewardship

Database or Data warehouse

Dynamic Approaches

- Adaptive query processing
- Personalization

Corrective Approaches

- Comparison to the real-world
- Data consolidation based on several sources or a reference source
- Cleaning, ETL
- Imputation of missing values

Main Challenges

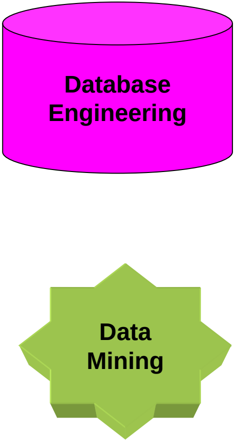
- **Methodological Level**
  - Unification and standardization
  - Benchmarks
- **Information System Engineering Level**
  - Design and architecture patterns for data quality control
- **Languages Level (DDL and DML)**
  - Declaration and integrated management of data and meta-data
  - Development and optimization of extended query languages
- **Algorithmic Level**
  - High dimensionality and volumetry of data and metadata
  - Data and metadata indexation
  - Optimization of statistical metadata computing
  - Dynamic awareness of data quality in the data processing

Proposed Approach

Mutual contributions of two fields

Axis 1 Using data mining techniques for data quality evaluation

Axis 2 Exploiting data quality metadata for evaluating and validating the quality of discovered knowledge and data mining results for decisional purposes

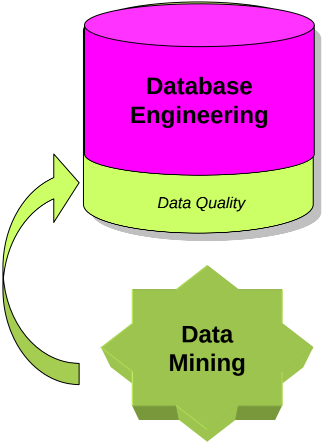


Proposed Approach

Mutual contributions of two fields

Axis 1 Using data mining techniques for data quality evaluation

Axis 2 Exploiting data quality metadata for evaluating and validating the quality of discovered knowledge and data mining results for decisional purposes

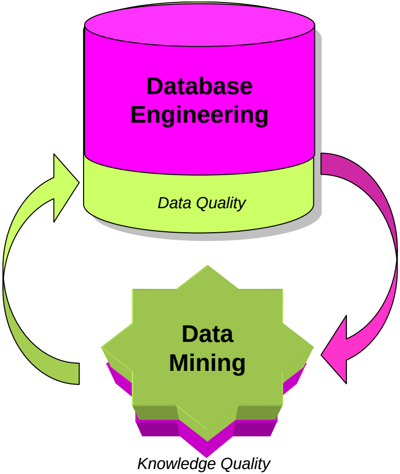


Proposed Approach

Mutual contributions of two fields

Axis 1 Using data mining techniques for data quality evaluation

Axis 2 Exploiting data quality metadata for evaluating and validating the quality of discovered knowledge and data mining results for decisional purposes



Axis 1: Quality-Awareness in Data management

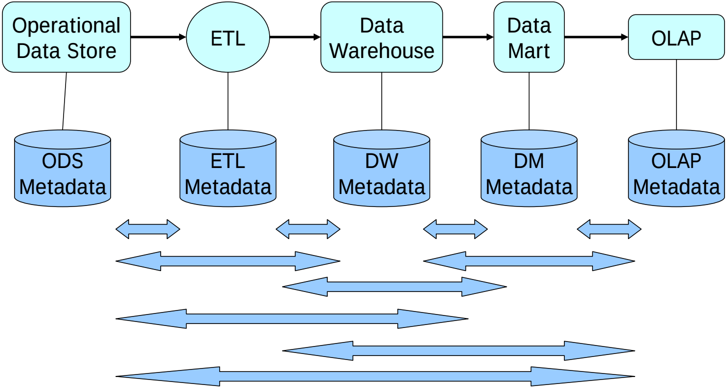
Objective: Computing and management of metadata describing measurable factors of data quality

Contributions:

1. Modeling metadata and joint management of data and metadata
2. Using and adapting statistical methods and data mining techniques for detecting patterns of anomalies on data
3. Extension of a query language for manipulating data quality metadata in the query processing

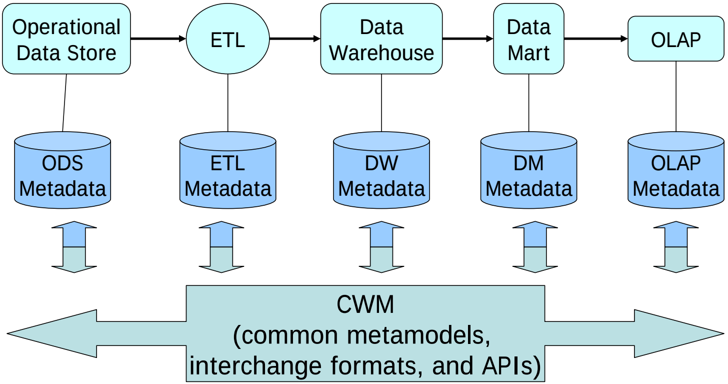
Extension of *Common Warehouse Metamodel* (OMG)

Problem of metadata integration: $\frac{n \times (n-1)}{2}$ exchanges

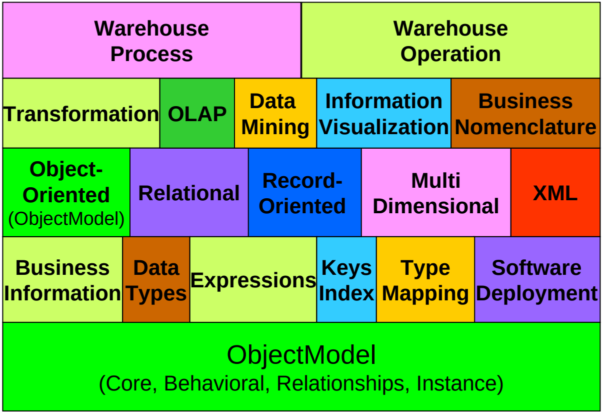


Extension of *Common Warehouse Metamodel* (OMG)

n CWM wrappers for metadata integration



CWM Packages



Management: **Warehouse Process**, **Warehouse Operation**

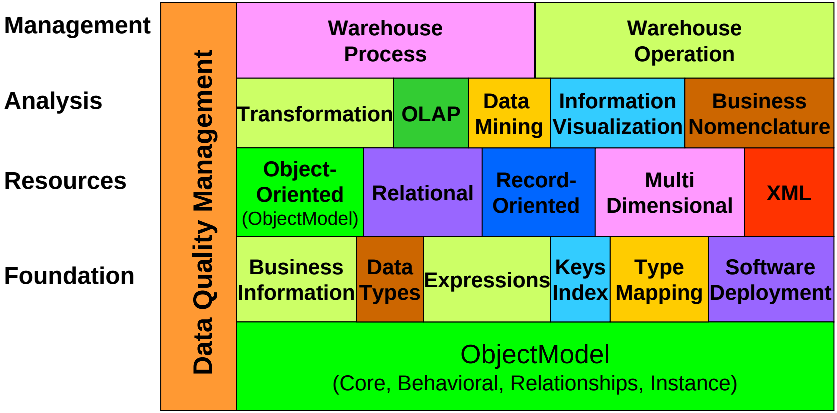
Analysis: **Transformation**, **OLAP**, **Data Mining**, **Information Visualization**, **Business Nomenclature**

Resources: **Object-Oriented** (ObjectModel), **Relational**, **Record-Oriented**, **Multi Dimensional**, **XML**

Foundation: **Business Information**, **Data Types**, **Expressions**, **Keys Index**, **Type Mapping**, **Software Deployment**

ObjectModel
(Core, Behavioral, Relationships, Instance)

CWM Packages



Management

Analysis

Resources

Foundation

Data Quality Management

Warehouse Process

Warehouse Operation

Transformation

OLAP

Data Mining

Information Visualization

Business Nomenclature

Object-Oriented (ObjectModel)

Relational

Record-Oriented

Multi Dimensional

XML

Business Information

Data Types

Expressions

Keys Index

Type Mapping

Software Deployment

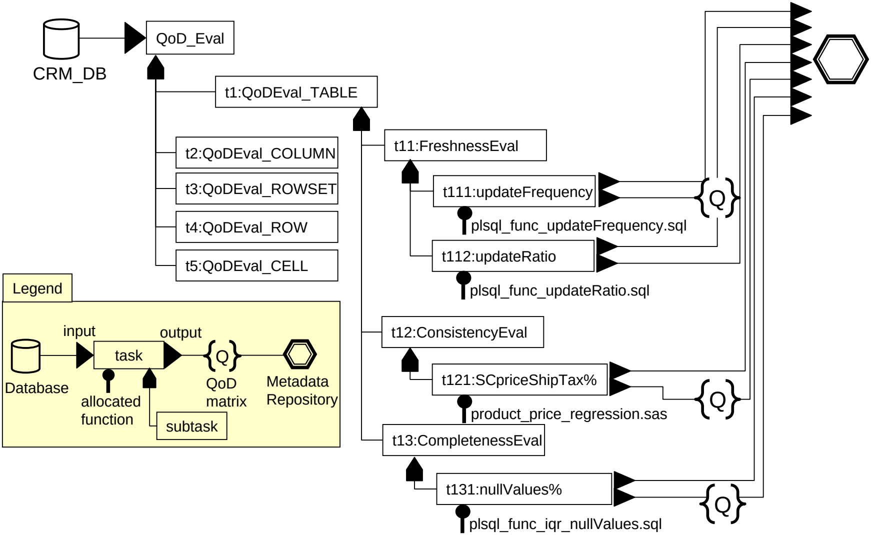
ObjectModel

(Core, Behavioral, Relationships, Instance)

Computing metadata with analytic functions

- Collect and define the functions useful for measuring data quality factors at different levels of granularity
  - **I:** Profiling functions
  - **II:** Constraint-based functions including statistical constraints
  - **III:** Synopses functions with sketches, histograms, and sampling techniques
  - **IV:** Mining functions
- Composition of functions in *analytic workflows*
- Storage and indexing of metadata

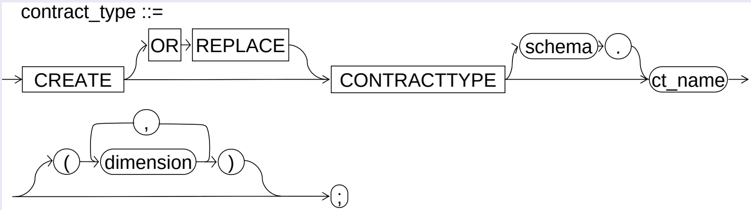
Example of an analytic workflow



Extension of SQL-like query language

Before querying

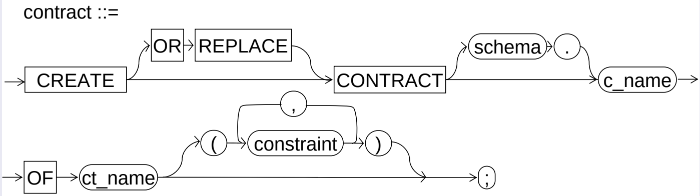
1. Creation of contract types composed of quality dimensions associated to one or more granularity levels
2. Creation of contracts with specified constraints on each dimension



Extension of SQL-like query language

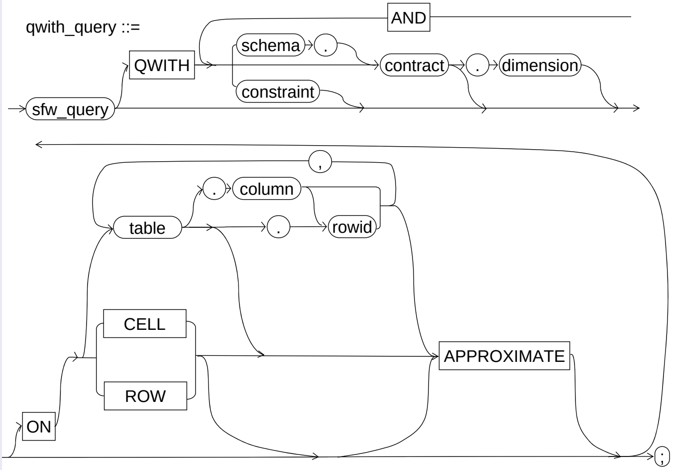
Before querying

1. Creation of contract types composed of quality dimensions associated to one or more granularity levels
2. Creation of contracts with specified constraints on each dimension



Extension of SQL-like query language

Quality-Constrained Query with contrats



Examples

Creation of Contract Types

```
CREATE CONTRACTTYPE FRESHNESS(
   timeliness FLOAT ON CELL,ROW BY FUNCTION func_timeliness
        IS LANGUAGE JAVA NAME './XQLib/func_timeliness.java');
CREATE CONTRACTTYPE COMPLETENESS(
   nullValues% FLOAT ON ROW, TABLE BY FUNCTION plsql_nullValues%);
CREATE CONTRACTTYPE CONSISTENCY(
   SCprice FLOAT ON PRODUCT BY FUNCTION price_regression
    IS LANGUAGE SAS NAME './XQLib/price_regression.sas');
```

Creation of Contracts

```
CREATE CONTRACT fresh OF FRESHNESS(timeliness > .50);
CREATE CONTRACT complete OF COMPLETENESS(nullValues% <= .80);
CREATE CONTRACT consistent OF CONSISTENCY(SCprice< .05);
```

Extended Query

```
SELECT PROD_ID, CUST_ID, FN, LN
FROM CUSTOMER C, PRODUCT P WHERE P.CUST_ID=C.CUST_ID
QWITH fresh ON CELL AND complete ON ROW AND consistent;
```

Contributions

- Modeling data quality metadata
- Development of a library of functions dedicated to data quality evaluation
- Design of *analytic workflows* for evaluating and controlling data quality
- Metadata manipulation with an extended query language

Perspectives

- Extended query optimization: approximation and constraint relaxation
- Extension of the library and development of a tool for helping the design of analytic workflows

Axis 2: Quality-Awareness in Data Mining

- Evaluate the cost of data non-quality on the knowledge discovered from rule mining technique
- Propose a probabilistic decision model based on data quality metadata
- Ensure the quality of discovered and legitimately interesting knowledge for decision-making

Interestingness Measures

Given the association rule R: $A \rightarrow B$ with A and B, two itemsets such as: $A \cap B = \emptyset$, the main interestingness measures are:

Support:	$\frac{N_A - N_{A\bar{B}}}{N}$
Confidence:	$1 - \frac{N_{A\bar{B}}}{N_A}$

- A rule is said to be valid if its confidence is greater than a predefined confidence threshold σ_C, and its support is greater than a predefined support threshold σ_S.
- A rule is said to be exact if its confidence is 1, otherwise the rule is partial.

Limit: Ignorance of quality metadata of the analyzed data

Measuring the Rule Quality

The quality of the association rule R: $A \rightarrow B$ is defined as:

$$Q(R) = \begin{pmatrix} q_1(R) \\ q_2(R) \\ \dots \\ q_k(R) \end{pmatrix} = \begin{pmatrix} q_1(A) \otimes_1 q_1(B) \\ q_2(A) \otimes_2 q_2(B) \\ \dots \\ q_k(A) \otimes_k q_k(B) \end{pmatrix}$$

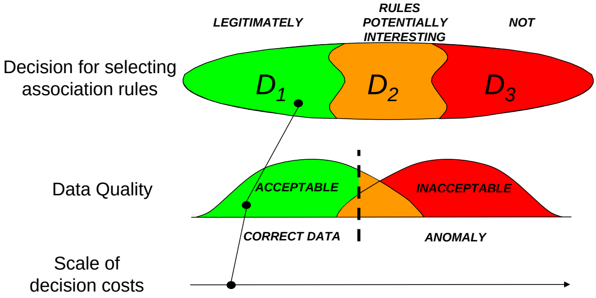
with $q_j(A)$ and $q_j(B)$, the measures associated to quality dimension j and computed for A and B composing the rule R and

$\otimes_j$ a particular fusion function per dimension, for example:

j	*Dimension*	*Fusion Function* $\otimes_j$
1	Freshness	$\min[q_1(A), q_1(B)]$
2	Consistency	$q_2(A) \cdot q_2(B)$
3	Completeness	$q_3(A) + q_3(B) - q_3(A) \cdot q_3(B)$

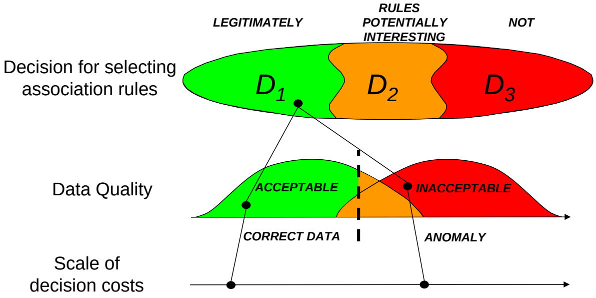
Objectives

1. Evaluate the average cost of a decision for selecting an association rule only based on interestingness measures ignoring initial data quality
2. Minimize the average cost with considering the probabilities that metadata reflect the actual data quality status (no anomaly detection problem).



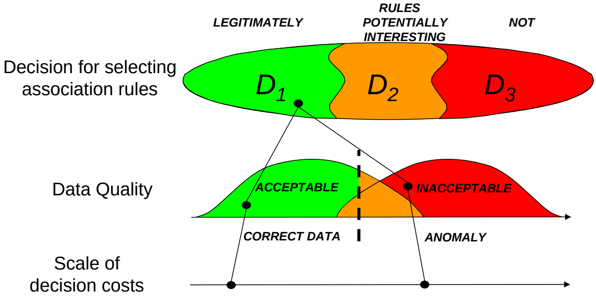
Objectives

1. Evaluate the average cost of a decision for selecting an association rule only based on interestingness measures ignoring initial data quality
2. Minimize the average cost with considering the probabilities that metadata reflect the actual data quality status (no anomaly detection problem).



Objectives

1. Evaluate the average cost of a decision for selecting an association rule only based on interestingness measures ignoring initial data quality
2. Minimize the average cost with considering the probabilities that metadata reflect the actual data quality status (no anomaly detection problem).



Experiments

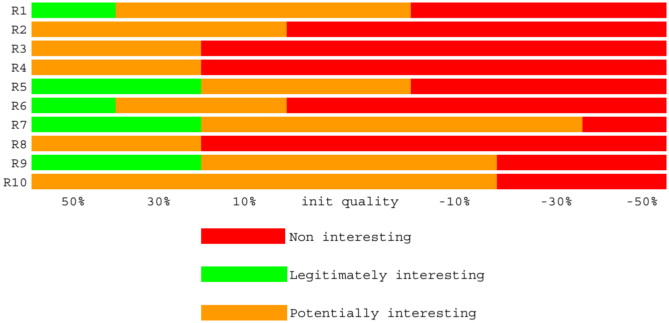
Experiments on KDD Cup-98 datasets:

- Extraction of the top ten association rules
- Generation of synthetic metadata describing data quality
- Evaluation of the rules as legitimately, potentially or non interesting rules
- Variations of data quality
- Cost analysis of data quality-blind decision based on selected rules based on acceptable vs unacceptable data quality

Experiments

Interesting Results:

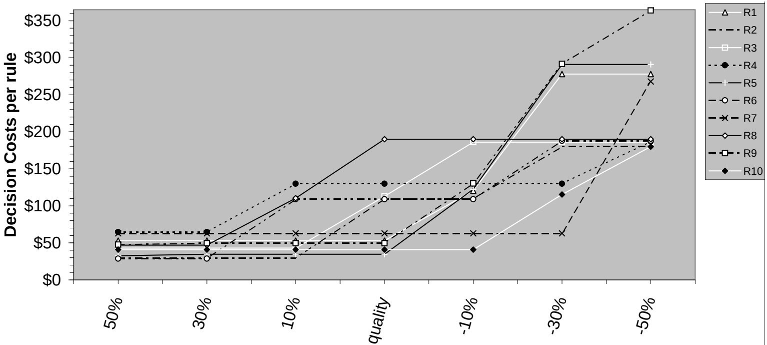
- Best rules are not always legitimately interesting : interestingness measures are not self-sufficient.
- Data quality deterioration implies significative decision cost increases.



Experiments

Interesting Results:

- Best rules are not always legitimately interesting : interestingness measures are not self-sufficient.
- Data quality deterioration implies significative decision cost increases.



Contributions and perspectives

- Exploitation of data quality metadata for:
  - Evaluating the quality of association rules and validation
  - Post-filtering association rules
- Retro-analysis and targeted corrective actions on data used for exploratory mining and decision-making
- Application to other mining techniques

Plan

3 **Applications**

- Integration of Genomic and Biomedical Data
- CRM Data Mediation
- Telecom Data Stream Monitoring

Project and collaboration with INSERM U522

Context

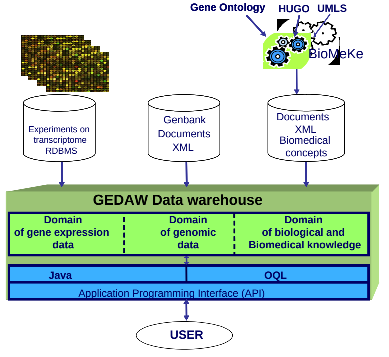
Collect all genomic and biomedical information and knowledge available in public databanks describing genes involved in liver pathologies

Contributions

- Modeling of the genomic domain
- Design of a specific ETL process (XML → OODW)
- Evaluation of biomedical data quality in the DW
- Development of tools for data warehouse exploration with browsing, querying, and profiling functionalities useful for biologists

Architecture: Data Integration System

- Extraction and cleaning of XML data from the main public databanks (GenBank, SwissProt)
- Integration into the object-oriented data warehouse: GEDAW (*Gene Expression DAta Warehouse*)


Gene Ontology
HUGO
UMLS
BioMeKe
Experiments on transcriptome RDBMS
Genbank Documents XML
Documents XML Biomedical concepts
GEDAW Data warehouse
Domain of gene expression data
Domain of genomic data
Domain of biological and Biomedical knowledge
Java
OQL
Application Programming Interface (API)
USER

Customer relationship Data

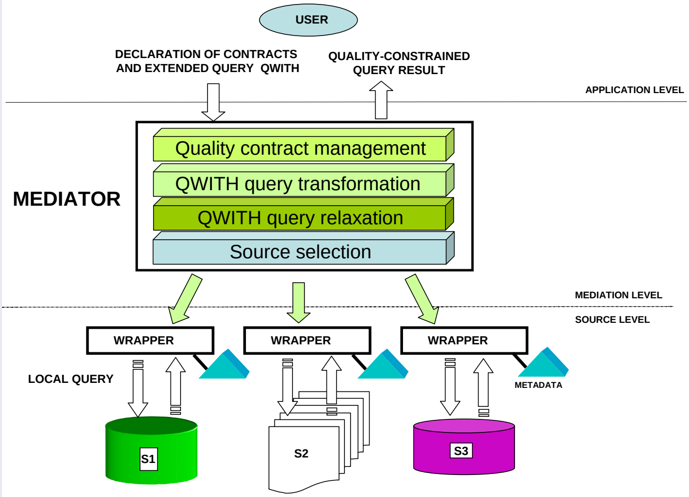
Context

Extended query processing including constraints on data quality in a mediation environment

Contributions

- Declaration and propagation of quality contracts
- Query language extended with `QWITH` operator
- Transformation of global extended queries (`SFW-QWITH`) into local extended queries
- Source selection depending on sources' ability to answer the query and satisfy the constraints on data quality
- Negotiation and relaxation of data quality constraints

Architecture: Data Mediation System



Telecommunication Data

Context

Prospective work for Genielog/SFR-Cegetel companies

Problem Statement

- On-line analysis and processing
- Stringent Constraints
- Approximation and widowing requirements

Contributions

- Study of data mining techniques for evaluating stream data quality
- Specification of first analytic workflows for stream data quality control

Plan

4 **Conclusions**

- Contributions
- Perspectives

Main Contributions

Data quality awareness in data management

- Modeling data quality metadata
- Specification of analytic functions for metadata generation
- Extension of a query language for declaration and manipulation of constraints on data quality

Data quality awareness in rule mining

- Exploitation of metadata for evaluating the quality of association rules
- Decision model for filtering legitimately interesting association rules with data quality awareness

Main Contributions

Data quality awareness in data management

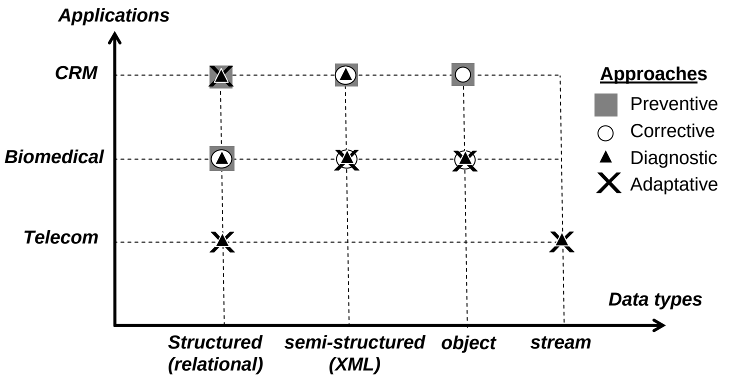
- Modeling data quality metadata
- Specification of analytic functions for metadata generation
- Extension of a query language for declaration and manipulation of constraints on data quality

Data quality awareness in rule mining

- Exploitation of metadata for evaluating the quality of association rules
- Decision model for filtering legitimately interesting association rules with data quality awareness

Applications

- Various domains
- Different data types
- Different approaches and architecture types



Research Directions

Short Term

- Optimizing extended queries
- Designing patterns of analytic workflows dedicated to the evaluation of data quality
- Studying the sensibility of clustering and mining techniques face to combined data quality problems

Research Directions

Mid Term

- Analysis of interdependencies between data quality dimensions: **QUADRIS project**
- Design of introspective data management systems: **mobility project funded by the European Commission**, 2 years in D. Srivastava's Team at AT&T Labs Research, New Jersey, USA

Research Directions

Mid Term

- Analysis of interdependencies between data quality dimensions: **QUADRIS project**
- Design of introspective data management systems: **mobility project funded by the European Commission**, 2 years in D. Srivastava's Team at AT&T Labs Research, New Jersey, USA

Long Term

Widening the coverage of my contributions to Data Quality Research to:

- Other application domains
- Much larger data volumes (3 billions of records)

Thanks !