Information Quality Measurement in Data Integration Schemas

10/2/2007

Maria da Conceição Moraes Batista &
Ana Carolina Salgado
Centro de Informática, UFPE
Recife - Brazil

Motivation

- Information Quality (IQ) has become a critical aspect in organizations and research areas
- IQ is a multidimensional aspect:
  - Consistency,
  - Availability,
  - Response Time,
  - Minimality,
  - Completeness,
  - …

In Data Integration Systems

- Data is spread over multiple, distributed and heterogeneous sources
- Our data integration system:
  - Mediator-based architecture
  - Global-as-view (GAV) approach to provide a unified view of several data sources: the *integrated schema*

Quality in Data Integration Systems

- The query execution quality is an essential feature

- Not so much is known about incorporating IQ aspects into data integration processes:
  - Query results integration,
  - Schema maintenance,
  - Mediator evolution,
  - ...

- Our goal:
  - Quality of query execution
- Our hypothesis:
  - The construction of good schemas with high quality scores will improve query execution
- Our proposal:
  - IQ analysis to address schema maintenance, specially the integrated schema
  - IQ criteria for data integration aspects
  - The specification of schema IQ criteria – *minimality*, *completeness* and *type consistency*

Outline

- Data Integration IQ Criteria
- Schema IQ Criteria Specification
  - Minimality
  - Schema Completeness
  - Type Consistency
- Schema Quality Improvement
- Conclusions & Ongoing Work

Data Integration IQ Criteria

Schema Quality

- The user submits queries to the *integrated schema*
  - A set of views over a number of data sources
- The data integration system reformulates a user query into queries that refers directly to schemas on the sources.
  - Schema mappings: correspondences between data sources and integrated schema elements

Data Integration IQ Criteria
Classification

- Three classes of components: *data*, *schemas* and *data sources*

Data Integration Element	IQ Criteria
Data Sources	Reputation, Verifiability, Availability, Response Time
Schemas	Schema Completeness, Minimality, Type Consistency
Data	Data Completeness, Timeliness, Accuracy

Schema IQ Criteria

- ***Schema Completeness***
  - Percentage of real-world objects modeled in the integrated schema that can be found in the sources
- ***Minimality***
  - The extent in which the schema is compactly modeled without redundancies.
  - The more minimal the integrated schema is, the least redundancies it contains, and, consequently, the more efficient the query execution

Schema IQ Criteria

- ***Type Consistency***
  - The extent in which the attributes corresponding to the same real world concept are represented with the same data type across all schemas

IQ Manager

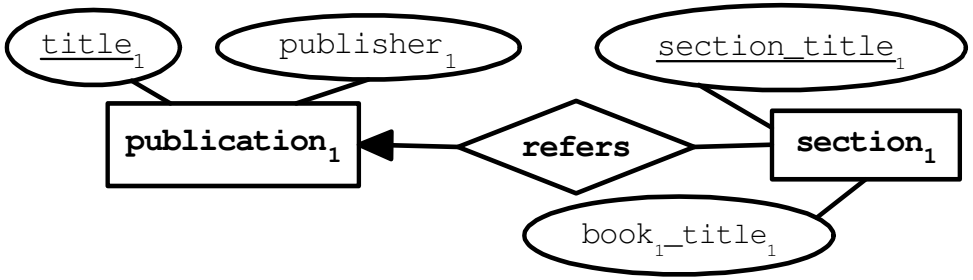
- A module of the data integration system
- It executes the IQ criteria analysis, assessment and adjustments over the schema to improve its IQ scores

Schema Representation

The X-Entity Model

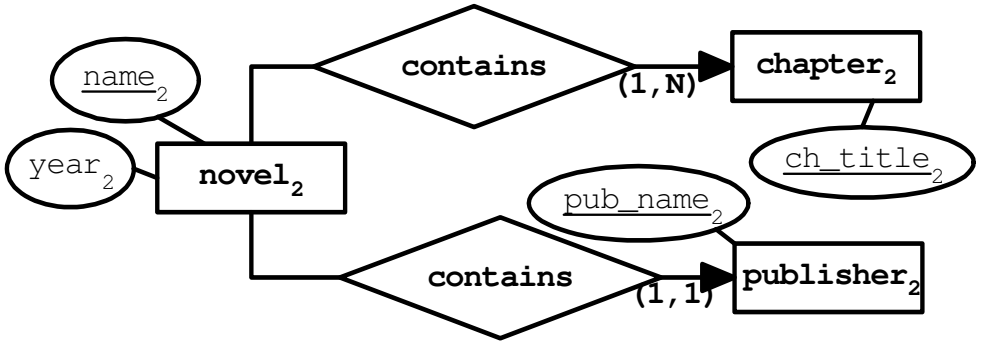
- E-R extension for XML data
- X-Entity Components
  - Entity = XML Elements
  - Relationships = XML relationships
    - Contains
    - Refers
  - Attributes

Representation



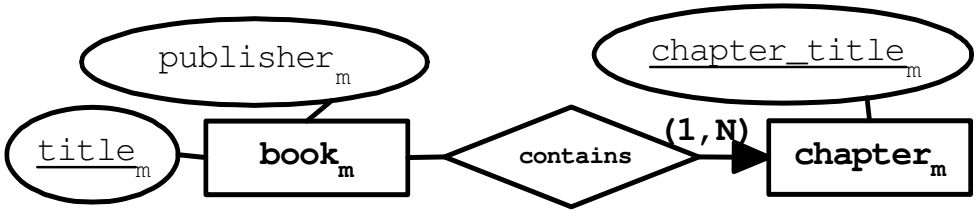
Source Schema S_1 =

```
({publication1({title1,publisher1},{}),
 section1({section_title1,book_title1},
          {ref_section1_publication1})},
{ref_section1_publication1
 (section1,publication1,{book_title1},{title1})})
```



Source Schema S_2 =

```
({novel2({name2,year2},
 {novel2_chapter2,novel2_publisher2}),
  chapter2({ch_title2},{}),
  publisher2({pub_name2},{})},
 {novel2_chapter2(novel2, chapter2,(1,N)),
  novel2_publisher2(novel2, publisher2, (1,1))}
```



Integrated Schema S_{med} =

```
({bookm({titlem,publisherm},{bookm_chaptermm}),
  chaptermm({chapter_titlem},{})},
 {bookm_chaptermm(bookm,chaptermm,(1,N))})
```

Schema Mappings

- Correspondences between X-Entities elements of source and integrated schemas representing the same real world concept (*semantically equivalent)*

$MP_1: book_m \equiv publication_1$

$MP_2: book_m.title_m \equiv publication_1.title_1$

$MP_3: book_m.publisher_m \equiv publication_1.publisher_1$

$MP_4: chapter_m \equiv section_1$

$MP_5: chapter_m.chapter\_title_m \equiv section_1.section\_title_1$

$MP_6: book_m.book_m\_chapter_m.chapter_m \equiv (section_1.section\_ref\_ publication_1.publication_1)^{-1}$

$MP_7: book_m \equiv novel_2$

$MP_8: book_m.title_m \equiv novel_2.name_2$

$MP_9: chapter_m \equiv chapter_2$

$MP_{10}: book_m.book_m\_chapter_m.chapter_m \equiv novel_2.\ novel_2\_chapter_2.chapter_2$

$MP_{11}: chapter_m.chapter\_title_m \equiv chapter_2.ch\_title_2$

$MP_{12}: book_m.publisher_m \equiv novel_2.\ novel_2\_publisher_2.publisher_2.pub\_name_2$

Some Definitions

Redundancy

- Attribute Redundancy:
  - An attribute A_1 in schema S_m is considered redundant, i.e.

    $Red(A_1, S_m) = 1$

    if $\exists$ A_2 in schema S_m and $A_1 \equiv A_2$

Redundancy

- Entity Redundancy:
  - The number of redundant attributes defines the entity redundancy:

$$Red(E_k, S_m) = \frac{\sum_{i=1}^{a_k'} Red(A_{ki}, E_k)}{a_k}$$

  - where $\sum_{i=1}^{a_k} Red(A_{ki}, E_k)$ is the total number of redundant attributes in entity E

Redundancy

- Relationship Redundancy
  - A relationship between two entities is redundant if there are other semantically equivalent relationships which paths are connecting the same two entities

Minimality

- A schema is minimal if all relevant domain concepts are described only once
- The minimality of a schema is the degree of absence of redundant elements
- A minimal schema will improve the effectiveness of operations and queries over it

Minimality

- The redundancy of a schema in a data integration system is measured by the sum of all redundancy values: entities redundancy (ER) and relationships redundancy (RR)

- The schema minimality is measured by the formula:

$$Mi_{S_m} = 1-[ER(S_m)+RR(S_m)]$$

Schema Completeness

- The schema completeness is the percentage of domain concepts represented in the integrated schema when related to the concepts represented in all data source schemas

- Example:
  - A data integration system with 10 distinct domain concepts
  - Described by entities and relationships in all data sources' published schemas
  - If the integrated schema includes 8 of these concepts, then the integrated schema is 80% complete related to the current set of data sources

Schema Completeness

- The overall schema completeness degree in a given schema $S_x \in$ **Đ** is obtained by the average:
  - $SC(S_x) = \dfrac{\sigma_{S_x}}{\sigma_{Đ}}$
  - S_x can be either a data source schema or the integrated schema;
  - σ_{S_x} is the number of distinct concepts in the schema S_x;
  - $\sigma_{Đ}$ is the is the number of distinct concepts contained in all the schemas of the data integration system **Đ**

Type Consistency

- When an integrated schema management system experiences problems with consistency, the same information is stored with more than one data type

- How to fix:
  - To determine which alternative data type is preferable (standard)
  - A schema element is consistent if it adheres to the standard data type

Type Consistency

- The type consistency metric is based in:
  - The number of semantically equivalent attributes in schema that adhere to the standard data type defined for the attribute

- Attribute Type Consistency
  - A given attribute A_{pj} is ***consistent*** i.e.

    $$\mathbf{Con(A_{pj}, S_p) = 1}$$

    if every semantically equivalent attribute to A_{pj} appears in another entity or even in the same entity with the standard data type of A_{pj}

Type Consistency

- The overall schema type consistency score in a given data integration system ($Con(S_m,Đ)$) is obtained by:
  - $Con(S_m,Đ) = \frac{\sum_{k=1}^{n_m}\sum_{j=1}^{a_k} Con(A_{kj},Đ)}{\sum_{k=1}^{n_m} a_k}$, where
  - $\sum_{k=1}^{n_m}\sum_{j=1}^{a_k} Con(A_{kj},Đ)$ is the total number of consistent attributes in **Đ**;
  - n_m is the total number of entities in the schema **Đ**
  - a_k is the number of attributes of the entity E_k

Schema Quality Improvement

Minimality

Minimality Improvement

- In order to improve minimality scores, redundant elements must be removed from the schema, until the value of minimality equal to 1 (no redundancies) is achieved

Algorithm for Schema Minimality Improvement

1	Calculate minimality score and if minimality = 1, then stop;
2	Search for fully redundant entities in S_m;
3	If there are fully redundant entities then eliminate the redundant entities from S_m;
4	Search for redundant relationships in S_m;
5	If there are redundant relationships then eliminate the redundant relationships from S_m;
6	Search for redundant attributes in S_m;
7	If there are redundant attributes then eliminate the redundant attributes from S_m;
8	Go to Step 1

Redundant Entity Elimination

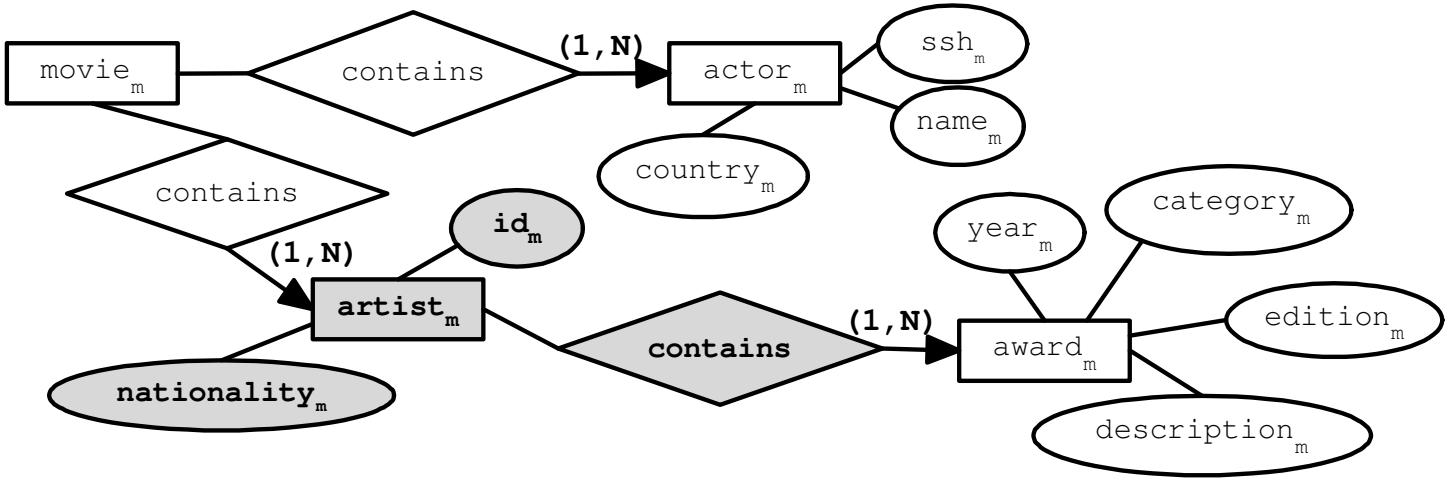
- When removing a redundant entity E_1 ($E_1 \equiv E_2$), the *IQ Manager* transfers the relationships of E_1 to the remaining equivalent entity E_2.

- Three different situations may occur when moving a relationship R_x, $R_x \in E_1$:
  - If $R_x \in E_2$ then R_x is deleted because it is no longer necessary;
  - If $R_x \notin E_2$ but $\exists R_y$, $R_y \in E_2$ such as $R_x \equiv R_y$ then R_x is deleted;
  - If $R_x \notin E_2$ and there is no R_y, $R_y \in E_2$ such as $R_x \equiv R_y$, then R_x is connected to E_2.

Redundant Relationships & Attributes Elimination

- Elimination of redundant relationships by deleting relationships identified as redundant
- Elimination of remaining redundant attributes in schema by deletion
- *IQ Manager* recalculates and analyzes minimality scores in order to determine if the desired IQ is accomplished

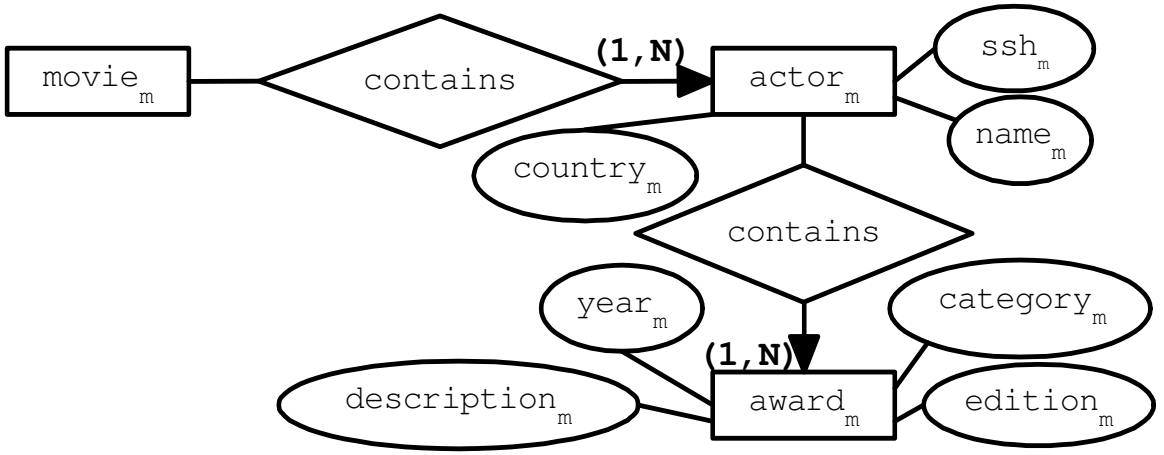
Example



$artist_m \equiv actor_m$

$id_m \equiv ssh_m$ **=> Red(artist$_m$, S$_m$) = 1**

$nationality_m \equiv country_m$



Implementation Issues

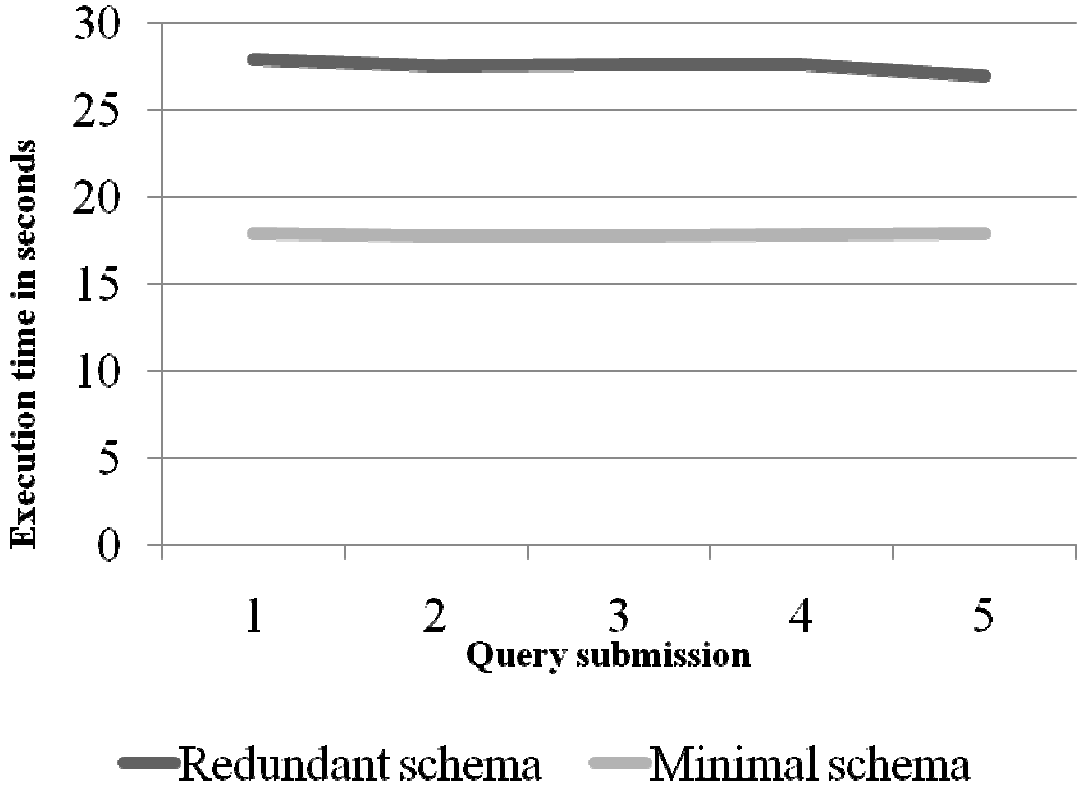
- IQ Manager is a Java module of *Integra* system
  - MySQL and PostgreSQL – data sources.
  - XML and XML Schema

- Experiment steps:

i. Queries were submitted over an integrated schema with 26% of redundant elements;
ii. Redundancy elimination algorithm executed generating a minimal schema (100% of minimality);
iii. Same queries of step (i) were re-executed.

Initial Experimental Results

- Query performance was improved in an average of 35%.



Conclusions

- We propose a quality approach that serves to analyze and improve the integrated schema definition and query execution
- Contributions:
  - Specification of IQ criteria assessment methods for the maintenance of high quality integrated schemas
  - Algorithm to improve the schemas' minimality scores.
  - The *IQ Manager* module to proceed with all schemas IQ analysis and also the execution of improvement actions by eliminating the redundant items

Ongoing Work

- Specification and implementation of algorithms to evaluate others IQ criteria
  - so far we are working in completeness and type consistency algorithms
- Experimentation of schema IQ improvement actions for each one

Information Quality Measurement in Data Integration Schemas

Maria da Conceição Moraes Batista (mcmb@cin.ufpe.br)
Ana Carolina Salgado (acs@cin.ufpe.br)
Centro de Informática, UFPE
Brazil