Research Issues
CIn/UFPE

Ana Carolina Salgado

Summary

- Schema Quality
- Name Resolution
- SPEED Project

Schema Quality Analysis in a Data Integration System

Maria da Conceição Moraes Batista
Ana Carolina Salgado

Contents

- Introduction
- Schema quality
  - Minimality
  - Type consistency
  - Schema completeness
- Main contributions

This work

- Our goal:
  - Quality of query execution in data integration systems
- Our hypothesis:
  - The construction of good schemas with high IQ scores will improve query execution
- Our proposal:
  - The specification of schema IQ criteria – *minimality*, *type consistency* and *schema completeness*
  - IQ analysis of the integrated schema in a data integration system

Schema issues in data integration

Schemas in Data Integration

- 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

Schema Representation

- The **X-Entity** Model
  - E-R extension for XML data
- X-Entity:
  - **Entity** = XML Elements
  - **Relationships** = XML relationships
    - **Contains:** a XML element contains other XML element
    - **Refers:** a XML element refers other XML element
  - **Attributes**
- The same real world concept: ***semantical equivalence*** (≡) already defined!!!

Minimality

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
- A schema is *minimal* if all relevant domain concepts are described only once
  - No redundant elements ➡ **Redundancy evaluation**

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 $A_1 \equiv A_2$

  - Otherwise, the attribute A_1 in schema S_m is considered **non redundant** , i.e.

    $Red(A_1, S_m) = 0$

Redundancy

- Entity Redundancy (ER)
  - The sum of attributes redundancy 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_K and a_k is the total number of attributes in E_k.

Redundancy

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

- The minimality concept is based in the redundancy evaluation

Minimality

- The schema redundancy is measured by the sum of all redundancy values: entity 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 Redundancy

Type Consistency

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
- A schema is ***type consistent*** if all its equivalent attributes are represented with the same data type
- The query becomes more efficient: less type conversions executed by the wrappers in composing results
- How to evaluate:
- 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*** in schema S_p i.e.,

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

    if every semantically equivalent attribute to A_{pj} appears in the schema S_p with the standard data type of attribute A_{pj}
  - Otherwise A_{pj} is ***inconsistent*** ,i.e., $\mathbf{Cst(A_{pj},S_p)= 0}$

Type Consistency

- The overall schema type consistency score in a given data integration system ($Cst(S_m,Đ)$) is obtained by:

$$Cst(S_m,Đ) = \frac{\sum_{k=1}^{n_m}\sum_{j=1}^{a_k} Cst(A_{kj},S_m)}{\sum_{k=1}^{n_m} a_k}$$, where

  - $\sum_{k=1}^{n_m}\sum_{j=1}^{a_k} Cst(A_{kj},S_m)$ is the sum of attribute consistency in S_m
  - n_m is the total number of entities in Đ
  - a_k is the number of attributes of the entity E_k

Schema Completeness

Schema Completeness

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

- The overall schema completeness degree in a schema S_x in a data integration system **Đ** is:
  - $SC(S_x) = \frac{\sigma_{S_x}}{\sigma_{Đ}}$, where
  - 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**

Contributions

- Specification of three relevant schema IQ criteria – *minimality*, *type consistency* and *schema completeness*
- Analysis of system schema elements according to the specified minimality and type consistency criteria of an integrated schema.
- Specification, implementation and tests of the IQ Manager in a data integration system with a real health care application

Thanks !!!