From Database To Datawarehouse: a Design Quality Evaluation

Maurizio Pighin and Lucio Ieronutti
IS&SE-Lab, Department of Computer Science
University of Udine (Italy)

Content

• Introduction and motivations
• Proposed Methodology
• Experimental evaluation
• Conclusion and future work
Introduction (1)

• Information systems allow organizations to collect a large number of transactional data
• Data warehousing provides architectures and tools to derive information at a level of abstraction suitable for supporting decision processes
• The quality of original data and the structure of original Databases (DBs) inevitably limits the quality of the decisions taken analysing the data
• The effectiveness of a Datawarehouse (DW) is strongly constrained by the quality of data selected for the analysis

Introduction (2)

• The DW designer has usually not a complete and global knowledge of the DB design and it is difficult to analyze the quality of the data of the original DB
• It is fundamental to select appropriate attributes to define measures or dimensions during the DW design process
• Our methodology can help the DW designer in
  • Select more appropriate elements
  • Evaluate the final quality of the DW
Data Quality

- Different definitions and metrics have been proposed in the literature (e.g., Wang and Redman); however, there is not a general agreement on data quality definitions.
- Concept typically related with data semantics, and then it is difficult to perform objective measurements.
- Evaluating data quality:
  - in Cooperative Information Systems (CIS), data replication is used for evaluating (and improving) data quality.
  - quality measurements through subjective metrics (e.g., questionnaires); this solution results more suitable for qualitative evaluations rather than quantitative ones.
  - techniques based on statistical analysis of data (e.g., data distribution).

Proposed methodology (1)

- Our goal is to propose a semantic independent methodology that can be effectively used to effectively support and evaluate the DW design choices.
- For such purposes, we have defined a set of metrics measuring different statistical and syntactical characteristics of original data and of DW structure.
- Each metric is designed with the aim of capturing a specific feature of data; quality information is derived by combining computed indexes.
- Proposed solution can be coupled with other procedural design approaches (e.g., Data Fact Model - DFM) for effectively driving the DW design choices.
Proposed methodology (2)

- During the DW creation
  - measure how much a DB table/DB attribute is suitable for extracting DW measures and/or dimensions
  - drive the selection of the DB attributes in the case of alternative choices (redundant information)
- At the end of the DW design
  - evaluate the final quality of DW and design choices

Proposed Indexes

- Given a DB element, we have defined a set of metrics measuring (statistical and syntactical) data features
- Indexes correspond to derived measurements
- We have considered the following types of DB elements:
  - Tables of a DB: for highlighting which tables contain more/less relevant data for decision process
  - Attributes of a table: for identifying which attributes of a considered table are more/less relevant data for decision process
- All indexes are normalized into the $[0, 1]$ interval
- These indexes are combined for obtaining quality indicators
Table Indexes (1)

• With this metrics, we aim at taking into account that different tables could play different roles and then result more/less suitable for extracting measures or dimensions

• The global indicators $S_{mj}$ and $S_{dj}$ indicating how much the table $t_j$ is suitable to extract measures and dimensions are obtained by combining derived indexes as follows:

$$S_{p,j} = \frac{\sum_{e=1}^{k} C_{p,e} \cdot M_e(t_j)}{k}$$

• $p = d$ or $m$ ($d =$ dimension, $m =$ measure)
• $e = 1, \ldots, k$ identifies the metric
• $j$ identifies the table
• $C_{p,e}$ corresponds to the table metric coefficient

Table Indexes (2)

• **Percentage of records**: percentage of records stored into the considered table with respect to the total number of records stored into the entire DB (length)
  • Time normalization
  • Transactional vs. descriptive data

• **Percentage of attributes**: percentage of table attributes with respect to the total number of DB attributes (width)

<table>
<thead>
<tr>
<th></th>
<th>$C_d$</th>
<th>$C_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_1$ - Percentage of records</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>$M_2$ - Percentage of attributes</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>

*List of coefficients*
Attribute Indexes (1)

- The global indicators $s_{m,i}$ and $s_{d,i}$ evaluating how much the attribute $a_i$ is suitable to be used respectively as measure and dimension are derived as follows:

$$s_{p,i} = \frac{\sum_{h=1}^{r} c_{p,h} \cdot m_h(a_i)}{r}$$

- $p = d$ or $m$ ($d = \text{dimension, } m = \text{measure}$)
- $h = 1, \ldots, r$ identifies the metric
- $i$ identifies the attribute
- $c_{p,h}$ corresponds to the attribute metric coefficient

Attribute Indexes (2)

- **Percentage of null values:** percentage of null values with respect to the total number of table records;
  - It limits all the other indexes
- **Degree of clusterization:** the extent in which the attribute assumes different values on the domain
  - High clusterization, for instance attribute of a business object useful for exploring data, but not useful for measuring elements
Attribute Indexes (3)

- **Dispersion of values**: how much data tends to spread over the domain (for identifying uniform distribution)
  - It traces if the distribution is uniform or not
  - For instance, price values which are useful for measures, are usually not uniformly distributed

- **Type of attribute**: string or date (0), short and integer (0.5), float and double (1)
  - Numerical attributes are more suitable for measures

- **Keys**: the attribute belongs to primary and/or duplicable keys?
  - Key-attributes are more suitable for dimensions

Attribute Indexes (4)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>(c_d)</th>
<th>(c_m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m_1) - Percentage of null values</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>(m_2) - Degree of clusterization</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>(m_3) - Dispersion of values</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>(m_4) - Type of attribute</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>(m_5) - Key</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>

List of coefficients
DW Metric (1)

• Each attribute is characterized by the couple of global indicators $G_{m,i,j}$ and $G_{d,i,j}$ indicating how much the attribute $a_i$ belonging to the table $t_j$ is suitable to be used respectively as measure and dimension.

• These global indicators are derived as follows:

$$G_{p,i,j} = S_{p,j} \cdot s_{p,j} \quad a_i \in t_j$$

- $p = d$ or $m$ ($d =$ dimension, $m =$ measure)
- $j$ identifies the table
- $i$ identifies the attribute
- $S_{p,j}$ corresponds to the table index
- $s_{p,j}$ corresponds to the attribute index

DW Metric (2)

• The global indicator $I(DW)$ measuring the final DW design quality is derived as follows:

$$I(DW) = \sum_{a_i \in A_d} G_{m,i,j} + \sum_{a_i \in A_m} G_{d,i,j}$$

- $n_d$ + $n_m$

- $A_d$ is the set of $n_d$ attributes chosen as DW dimensions
- $A_m$ is the set of $n_m$ attributes chosen as DW measures

• It is useful for instance to evaluate the quality of different DWs, i.e. they informative support to decision making.
Experimental Evaluation

- Tested on a subset (the selling part) of an enterprise commercial DB of a real world business system (22 tables, 528 attributes and millions of records)
- Procedure: (i) we asked to an expert to build a DW for analyzing selling information, selecting the attributes that are the most suitable to support decision processes, (ii) we tested our metrics evaluating the measured quality of selected attributes (iii) we evaluated the results
- Phases:
  - Calculated table metrics
  - Calculated attribute metrics
  - Final DW evaluation
  - Methodology Assessment

Phase 1: Table metrics

<table>
<thead>
<tr>
<th>Table</th>
<th>$S_d$</th>
<th>$S_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>xsr</td>
<td>0.5554</td>
<td>0.4446</td>
</tr>
<tr>
<td>intf</td>
<td>0.6884</td>
<td>0.3116</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>art</td>
<td>0.7322</td>
<td>0.2678</td>
</tr>
<tr>
<td>dii</td>
<td>0.7418</td>
<td>0.2582</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>smag</td>
<td>0.7491</td>
<td>0.2509</td>
</tr>
<tr>
<td>tbd</td>
<td>0.7494</td>
<td>0.2506</td>
</tr>
</tbody>
</table>

*xsr* stores selling information and then results suitable for extracting measures

*smag* stores information on products categories and then results suitable for extracting dimensions
Phase 2: Attribute metrics

Candidates for dimensions:
- product codes
- product categories
- ...

Candidates for measures:
- pricing information
- amount of products sold
- ...

Unsuitable from both dimension and measure point of view (high percentage of null values)

<table>
<thead>
<tr>
<th>Table</th>
<th>Attribute</th>
<th>$G_{ij}$</th>
<th>Rank$_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>liof</td>
<td>sigla_art</td>
<td>0.593</td>
<td>1</td>
</tr>
<tr>
<td>art</td>
<td>tipolog_art</td>
<td>0.553</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>0.166</td>
<td>348</td>
</tr>
<tr>
<td>xsr</td>
<td>valore</td>
<td>0.164</td>
<td>349</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>0.000</td>
<td>527</td>
</tr>
<tr>
<td>xsr</td>
<td>sconto_ex</td>
<td>0.000</td>
<td>528</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table</th>
<th>Attribute</th>
<th>$G_{mn}$</th>
<th>Rank$_{mn}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>xsr</td>
<td>qta</td>
<td>0.395</td>
<td>1</td>
</tr>
<tr>
<td>xsr</td>
<td>valore</td>
<td>0.394</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>0.118</td>
<td>336</td>
</tr>
<tr>
<td>art</td>
<td>tipolog_art</td>
<td>0.102</td>
<td>347</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>0.000</td>
<td>527</td>
</tr>
<tr>
<td>xsr</td>
<td>sconto_ex</td>
<td>0.000</td>
<td>528</td>
</tr>
</tbody>
</table>

Phase 3: DW Design Quality

- The DW built by the expert is characterized by 6 measures and 10 dimensions. We have evaluated
  - the DW measures and dimensions ranking
  - the final I(DW)=0.6292, due to some not very quality attribute

![Measures](chart1.png) ![Dimensions](chart2.png)
Phase 4: Methodology Assessment (1)

• The results of application of the methodology on a real Commercial DB are consistent with our expectations
• The principal tables and relative attributes useful to extract measure or dimensions from original DB are correctly highlighted
• They have different values (and so different ranking) depending
  • the quality of data contained (from informative point of view)
  • the quality of original DB design (for informative point of view)

Phase 4: Methodology Assessment (2)

• The DW designer, when chooses semantically a field, can directly evaluate the quality level of the data contained (or its eventually problems); for instance
  • the fields with “discount values” are not very informative measures because they have an high percentage of null values
  • the “district of the customer” is not very informative dimension, because the distribution of its values
Conclusion and Future Work (1)

- Methodology for both supporting the selection of DW dimensions and measures and evaluating the quality of taken design choices
- We plan to introduce additionally indexes to improve the accuracy of the measurements e.g. including information on data entropy and relations between tables
  - Now we are testing 13 indexes
    - 5 “table-metrics”, 6 “attribute metrics” and 2 “relation metrics”

Conclusion and Future Work (2)

- Since we have used unitary values, we intend to investigate if an accurate tuning of coefficients can further increase the accuracy of the measurements
- We are currently testing our metrics on 3 DBs (two corresponding to different instantiations of the same DB schema, while the third is characterized by a different schema but used to build the same DW)
- We are also evaluating the effectiveness of the methodology on different environments
  - Motorways crashes
  - University students’ examinations