Automatically Inferring ClassSheet Models from Spreadsheets

Jácome Cunha  João Saraiva  Martin Erwig
Universidade do Minho  Oregon State University

VL-HCC 2010 - Madrid
An Example

- This spreadsheet (SS) contains information about detergents
- Cells contain values and formulas
Models...

- Model-driven engineering uses models to help programmers.
- But in spreadsheets models are usually missing (e.g. legacy spreadsheets).
- It is difficult for end users to define the business model of a spreadsheet.
- Inferring them from data is a possible solution.
ClassSheet Models

- Object-oriented like specifications
- Contain spatial information, crucial for spreadsheets
- They are appropriate to specify the business logic of a spreadsheet
Functional Dependencies

- **Functional Dependency (FD):** $A \rightarrow B$
- We compute the business logic from the data, by inferring FDs
- They are the building blocks for constructing ClassSheet (CS) models
Some Examples of FDs

- com_code $\rightarrow$ upc, description
- size $\rightarrow$ upc
- nitem $\nrightarrow$ week
Inferring FDs

- We use a data mining algorithm to infer all FDs
- These algorithm produce too many FDs
- We use some heuristics to filter the “accidental” FDs
The Heuristics

**Label semantics**: usually keys are labeled “code” or “id”

**Label arrangement**: we prefer FDs respecting the order of columns

**Antecedent size**: small keys are preferable

**Ratio**: small antecedents and big consequents

**Single value columns**: columns always with the same value are too intrusive
Foreign Keys

• Set of columns referencing other columns in some table

• Necessary for finding relationships

<table>
<thead>
<tr>
<th>Candidate Key Att.</th>
<th>Foreign Key Att.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Schema</strong></td>
<td><strong>Attribute</strong></td>
</tr>
<tr>
<td>Price</td>
<td>price</td>
</tr>
<tr>
<td>Detergent</td>
<td>upc</td>
</tr>
<tr>
<td>StoreWeek</td>
<td>store</td>
</tr>
<tr>
<td>StoreWeek</td>
<td>week</td>
</tr>
<tr>
<td>StoreWeek</td>
<td>profit</td>
</tr>
</tbody>
</table>
Relational Intermediate Graph

- Graph containing the entities and the foreign keys
- Four entities all connected to “Sale”
Detecting Relationships

- “Sale” is a relationship because all its columns are FKs to other table

```
Sale

Price

StoreWeek

Detergent
```
Generating ClassSheets

- A \((A_1,\ldots,A_n)\), and default values \(d_{a_1},\ldots,d_{a_n}\)

- Two entities with a FK: the FK columns have references to the other table
Generate Relationships

\[ M \ (M_1, \ldots, M_r, M_{r+1}, \ldots, M_s) \]

\[ N \ (N_1, \ldots, N_t, N_{t+1}, \ldots, N_u) \]

\[ R \ (M_1, \ldots, M_r, N_1, \ldots, N_t, R_1, \ldots, R_x, R_{x+1}, \ldots, R_y) \]
Special Case I

M \((M_1, \ldots, M_r, M_{r+1}, \ldots, M_s)\)

N \((N_1, \ldots, N_t, N_{t+1}, \ldots, N_u)\)

R \((M_1, \ldots, M_r, N_1, \ldots, N_t, R_1, \ldots, R_x)\)

<table>
<thead>
<tr>
<th>A</th>
<th>Mkey</th>
<th>M_{1}</th>
<th>\ldots</th>
<th>M_{r}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nkey</th>
<th>M_{1} = M \cdot M_{1}</th>
<th>\ldots</th>
<th>M_{r} = M \cdot M_{r}</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>N_1 \ldots N_t</td>
<td></td>
<td>R_1 \ldots R_x</td>
</tr>
<tr>
<td>5</td>
<td>n_1 = N \cdot N_1 \ldots n_t = N \cdot N_t</td>
<td>r_1 = d_r_1 \ldots r_x = d_r_x</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>Sale</td>
<td>DishKey</td>
<td>Upc</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>StoreWeek</td>
<td>Move</td>
<td>store=&quot;&quot;</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>store=&quot;&quot;</td>
<td>week=&quot;&quot;</td>
<td>profit=move*</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Detergent</td>
<td>Upc</td>
<td>Com_code</td>
<td>Description</td>
<td>Size</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Price</td>
</tr>
<tr>
<td>15</td>
<td>Price</td>
</tr>
<tr>
<td>16</td>
<td>price=0</td>
</tr>
</tbody>
</table>
Special Case III

For cases with more than three classes:

• M and N should have small keys
• The number of empty cells created by the cell class should be minimal
Evaluation

• 27 spreadsheets, with a total of 121 worksheets
  (from the book *The Art of Modeling with Spreadsheets*)

• 66 out of the 121 contain formulas
Research Questions

**RQ1**: In how many cases is ClassSheet inference applicable?

**RQ2**: How many of the table and relationship structures that can be identified in the data can be successfully captured by ClassSheets inferred by our tool?

**RQ3**: In which cases does ClassSheet inference fail?
Results

Run the algorithms and visually evaluated the results:

- Inferred tables/relations: 176
- Failed: 13
- Bad results: 12
- Acceptable results: 27
- Good results: 124
Discussion

• We can produce ClassSheet models for 92% of the existing tables
• Of these, 76% are classified as good
• The failures occurred because no structure was found or FDs not in the data
• The results are encouraging, but could be improved by external information (more FDs or headers)
HaExcel

- Algorithms defined as reusable Haskell libraries
- Binding from Excel to Haskell
- Gumeric Haskell front-end
Conclusions

• We believe that models can improve spreadsheet usage (paper in preparation)

• But the usual lack of models is a problem

• We presented a technique to infer ClassSheet models from legacy spreadsheets

• The results are encouraging, but the technique could benefit from external info (more FDs or headers)