General Purpose Database Summarization
A web service architecture for on-line database summarization

Régis Saint-Paul (*speaker*),
Guillaume Raschia, Noureddine Mouaddib

LINA - Polytech’Nantes - INRIA
ATLAS-GRIM Group

VLDB Conference — Sept. 1st 2005
Table of Content

1 Introduction
   • Generalities
   • Related works

2 Summary model
   • Description space
   • Building the summaries

3 System architecture
   • Web service organization
   • Complexity and performances

4 Conclusion
Provide small versions of very large databases

- **Descriptive ability:**
  - scientific studies (epidemiology);
  - commercial and marketing studies (customer segmentation);
  - log analysis (connection/operation profile);
  - data obfuscation;
  - data personalization and filtering.

- **Data size reduction ability:**
  - approximate querying (hotel booking),
  - database browsing (image database),
  - storing rough view of the data on devices with low memory capacity (tourism GPS data).
Motivations

Provide small versions of very large databases

- **Descriptive ability**:  
  - scientific studies (epidemiology);  
  - **commercial and marketing studies** (customer segmentation);  
  - log analysis (connection/operation profile);  
  - data obfuscation;  
  - data personalization and filtering.

- **Data size reduction ability**:  
  - approximate querying (hotel booking),  
  - database browsing (image database),  
  - storing rough view of the data on devices with low memory capacity (tourism GPS data).
Motivations

Provide small versions of very large databases

- **Descriptive ability**:
  - scientific studies (epidemiology);
  - **commercial and marketing studies** (customer segmentation);
  - log analysis (connection/operation profile);
  - data obfuscation;
  - **data personalization and filtering**.

- **Data size reduction ability**:
  - approximate querying (hotel booking),
  - **database browsing** (image database),
  - storing rough view of the data on devices with low memory capacity (tourism GPS data).
What is a summary?

Definition
A summary is a concise representation of a set of structured data.
⇒ Semantic Compression

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ph.D. Student</td>
<td>1 000</td>
</tr>
<tr>
<td>Lecturer</td>
<td>2 000</td>
</tr>
<tr>
<td>Managing Director</td>
<td>8 500</td>
</tr>
<tr>
<td>Politician</td>
<td>xx xxx</td>
</tr>
</tbody>
</table>

**Tab.: Relation R**

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>Miserable</td>
</tr>
<tr>
<td>Executive</td>
<td>Enormous</td>
</tr>
</tbody>
</table>

**Tab.: Summary R**
Aggregate computation

- **Aggregate computation**
  
  *SDB, OLAP [Codd et al. 93], DataCubes [Gray et al. 93]*

- **Datacube summarization**
  
  *QuotientCube [Lakshmanan et al. 2002]*

**Limitations**

- Do not preserve the initial data schema;
- Subject oriented, has to be designed;
- Fixed and crisp granularity, threshold effect.
Clustering approaches for semantic compression

**Intuition**

Describe groups rather than individual observation.

- **Clustering** – *ItCompress* [Jagadish et al. 1999]
- **Bayesian network classifier** – *Spartan* [Babu et al. 2001]
- **Association rules** – *Fascicule* [Jagadish et al. 1999]

**Limitations**

- Classes shape depends on the selected criteria [Fasulo 1999];
- Single granularity of the compressed relation;
- Non-intuitive intentional description of classes.
Foundations of our approach

Intuition

Trying to reproduce the human learning mechanisms.

- **Formal concept analysis**
  

- **Conceptual clustering** – [Michalski et Stepp 1983]
  
  Unimem [Lebowitz 1986], Cobweb [Fisher 1987],
  Fuzz [Chen & Lu 1997]

Limitations

- Approaches were validated only on small data samples;
- Lack of maintenance capabilities.
Table of Content

1 Introduction
   • Generalities
   • Related works

2 Summary model
   • Description space
   • Building the summaries

3 System architecture
   • Web service organization
   • Complexity and performances

4 Conclusion
Possibilistic Data Representation

- Theoretical foundation:
  - Fuzzy-set theory (Zadeh, 1965) and Possibility theory (Zadeh 1978, Dubois & Prade 1985)
  
- Management of uncertain, incomplete and gradual information:
  “John’s age should *approximately* be *between 16 and 20*, but that’s *not sure*.”

- Possibility distribution
For each attribute $A$ with domain $D_A$, a set of Linguistic Labels is defined together with their membership function over $D_A$.

Example, on attribute INCOME:

\[
D_{\text{INCOME}} = [0, 200000] \\
D_{\text{INCOME}}^+ = \{\text{none, miserable, modest, ...} \}
\]
Summary representation space

Original tuple (raw data)

\[ t = \langle t.A_1, \ldots, t.A_k \rangle, \quad t \in \mathcal{R} \]

\[
\begin{align*}
\{t\} & \xrightarrow{D_A} \mathcal{R}(A_1, \ldots, A_k) = \prod_{i=1}^{k} D_{A_i} \\
\{z\} & \xrightarrow{\mathcal{F}(D_A^+)} \mathcal{R}^*(A_1, \ldots, A_k) = \prod_{i=1}^{k} \mathcal{F}(D_{A_i}^+) 
\end{align*}
\]

Summarized tuple

\[ z = \langle z.A_1, \ldots, z.A_k \rangle, \quad z \in \mathcal{R}^* \]
Summary model

A summary is a 3-uple $z = (I_z, R_z, E_z)$ with:

- $I_z$: the intentional content;
- $R_z$: the extensional content, subset of the relation $R$;
- $E_z$: a set of edges toward other summaries.

Example of a summary

<table>
<thead>
<tr>
<th>Label</th>
<th>satisfaction</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>intention $I_z$</td>
<td>1.83</td>
<td></td>
</tr>
<tr>
<td>OCCUPATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>employee</td>
<td>0.2</td>
<td>1.25</td>
</tr>
<tr>
<td>manager</td>
<td>1.0</td>
<td>0.33</td>
</tr>
<tr>
<td>managing director</td>
<td>0.7</td>
<td>0.25</td>
</tr>
<tr>
<td>INCOME</td>
<td></td>
<td></td>
</tr>
<tr>
<td>comfortable</td>
<td>1.0</td>
<td>1.50</td>
</tr>
<tr>
<td>high</td>
<td>1.0</td>
<td>0.33</td>
</tr>
<tr>
<td>extension $R_z$</td>
<td>{ $t_1$, $t_2$, $t_5$, $t_{13}$ }</td>
<td>4</td>
</tr>
</tbody>
</table>
Subsumption relation:

\[ z \subseteq z' \iff R_z \subseteq R_{z'} \]

Hierarchical organization:
- **root**: most general summary;
- **leaves**: most specific summaries.

The user-defined Background Knowledge fixes the finest level and, consequently, the maximal hierarchy size.
Algorithm outline

- hierarchical conceptual classification
- incremental process
- *top-down* approach
- selective local search

**Advantages**

summary freshness through incremental maintenance
linear time complexity w.r.t. the number of tuples

**Weaknesses**

sub-optimal model  
(order effect) (dynamic environment)  
(use of bidirectional learning operators)
Process overview

**Description space**

**Building the summaries**

### Summary model

**System architecture**

**Introduction**

---

### Process overview

[Diagram showing the process overview]

- **Original Database**
  - ID | Occupation | Income
  - t₁ | sax player  | 18 000
  - t₂ | unemployed  | 5 000
  - t₃ | pensioner   | 13 000
  - ...

- **Background Knowledge**

- **Candidate tuples**

- **Conceptual clustering**

- **Summary hierarchy**

  - R
    - R.1
      - R.1.1
    - R.2
    - R.3

  **Summary R.1.2**

  - **Intention**
    - Occupation | Income
    - 1/no occupation | 1/miserable
    - 1/artist

  **Extension:**
  - ct₁, ct₃[1]
The process looks for the learning operator which produces the highest quality child partition.

**Learning operators**
- affect,
- create,
- merge,
- split.

New candidate tuple $ct$
Local search

The process looks for the learning operator which produces the highest quality child partition.

Learning operators
- affect,
- create,
- merge,
- split.
The process looks for the learning operator which produces the highest quality child partition.

**Learning operators**
- affect,
- create,
- merge,
- split.
Local search

The process looks for the learning operator which produces the highest quality child partition.

Learning operators
- affect,
- create,
- merge,
- split.
Multi-granularity summary

The summary hierarchy presents many different precision levels.

\[ R_1^* \quad R_2^* \quad R_3^* \]

< \{ artiste, no occupation\}, \{miserable, none\} >

< \{ artiste, no occupation\}, miserable >
< no occupation, miserable >

< artiste, miserable >
< no occupation, none >

- The trade-off between size and concision can be chosen \textit{a-posteriori} depending on the user need;
- Analogy between the drill-down/roll-up operation on Datacube and the summary hierarchy navigation.
Table of Content

1 Introduction
   • Generalities
   • Related works

2 Summary model
   • Description space
   • Building the summaries

3 System architecture
   • Web service organization
   • Complexity and performances

4 Conclusion
- Message Oriented Application;
- Each document has autonomous specification (XSchema);
- Possibility to benefit from Message Oriented Middleware (MOM);
- Each service may be used separately or composed with others;
- Based on wide spread standards (W3C, ECMA et ISO).
Concept formation performed by autonomous “agents”

From translation service or other cooked data source

- Memory management optimized through specific pagination method;
- Process parallelization;
- Computation optimization through the use of a local cache with incremental upholding (contrast matrix).
Process performance evaluation

Tests based on 1990 US census data [UCI KDD Archive].

- 1 billion tuples;
- 14 attributes used for the summarization;
- 5 to 14 modalities per attributes (prepared).
Dynamic performances

Process performance is dependent only on the hierarchy size.

\[ \text{depth} = \log_{\text{width}}(\text{leaves}) \]
The marketing department of CIC (a french banking group) provided customer data:

- 33700 records;
- 70 attributes (10 of them used for the summary);
- Background Knowledge defined with the bank marketing experts;
- 3 to 8 linguistic descriptors used per attribute.
The number of leaves follows an asymptotic evolution;
The process tends toward a classification only regime.
Table of Content

1. Introduction
   - Generalities
   - Related works

2. Summary model
   - Description space
   - Building the summaries

3. System architecture
   - Web service organization
   - Complexity and performances

4. Conclusion
We presented:

- A general purpose multi-granularity summarization model:
  - an *adaptative* alternative to the *GROUP BY*;
  - simultaneous maintenance of *several compression levels*;
  - *robust* and *intuitive* classes thanks to human-like learning mechanism and uncertainty handling.

- The architecture of the system, which contributes to:
  - ease of *coupling* with DBMS (web services);
  - *performance* optimization and parallelization (use of autonomous agent);

- Validation of the system performance on a test database and a real-life one.
Web Site of SaintEtiQ

http://www.simulation.fr/seq

- Win32 prototype with test dataset available for download
- Process available online as web service
- References and documentation