Feature-Based Classification of Information Integration Systems

Dr. Susanne Busse
(in collaboration with Prof. Christoph Freytag, HU Berlin)

Database Research Group Meeting
University of Waterloo
20 July 2007

Research Topics @ CIS, TU Berlin

Continuous Software Engineering:
- I&C Infrastructures
- Model Based Software Development
- Methods and Tools

Main Focus of the whole Group

Heterogeneous Distributed Information Systems
- Integration Platforms / Interoperability
- Semantic Integration
- Information Logistics

Software-intensive technical systems
- Domain Engineering
- Embedded Systems

Outline

• Information Integration Systems (IIS) @ TUB
• How to Support the Design of IIS?
• Feature-Based Design: Reference Architectures
  • Feature-Based Classification of IIS
  • Feature Matching
• Features for ... – Some Ideas
• Conclusions
Information Integration Systems (IIS) @ CIS

Mediator-based IS

European Migration Network

Mediator

Web-Quelle

Central Information Integration Service

NCP

Local Metadata

Country C

Local Metadata

Country B

Local Metadata

Country A

Search Engines

Portals /

Requirements

Data Sources

Architecture

Computational-Independent Design

Platform-Independent Design

Platform-Specific Design

Implementation

Approach (1): Feature Modelling

Model about IIS Variants

Possible Features

Model

Schema Architecture

Software Architecture

Features / Characteristics

Objective

Support for engineers designing the integration layer of an IIS by

• specifying reusable elements – patterns – for IIS
• helping them to find and to use them
Outline

• Information Integration Systems (IIS) @ TUB
• How to Support the Design of IIS?
• Feature-Based Design: Reference Architectures
  • Feature-Based Classification of IIS
  • Feature Matching
• Features for ... – Some Ideas
• Conclusions

Classification of IIS

Classifications
[Hul 97] Discussion of Schema-based Systems
[PGB 00] Classification of Knowledge-based IIS

Descriptions of Types of IIS
[MSS+ 02], [Gar 04] Mediator-based IS
[HED+ 03], [HSB+ 05] Semantic Portals
[Rao 04], [ACG+ 01], [Bro 02] PDMS

Discussion of Particular Topics or Domains
[LC 03] Search Engines
[NS 06] Quality Integration in Bioinformatics

Approach (2): Identifying Patterns

Variants
Selected Features
What can be reused?
Features IR-IIS
Features MBIS
Matching
Variants

Approach (3): Applying Patterns

Model i
Select Model about Variants
Select / Remove Features
Transformation
Generated Model i+1
How to reuse?

2007, Susanne Busse
Technische Universität Berlin
Computation and Information Structures
### IIS Reference Architectures

**Mediator-based Information Systems**
- Mediator
- Wrapper

**Peer Database Management Systems**
- Peer
- Peer

**Semantic Portals** (mainly for Web Site Management)
- Portal

**Information Retrieval – Based Integration Systems**
- Index-based Query Processor

**Semantic Portals**
- (mainly for Web Site Management)

### Specifying IIS Reference Architectures

<table>
<thead>
<tr>
<th>Varianten IIS</th>
<th>MBIS</th>
<th>IR-IIS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Information Discovery Processes</strong></td>
<td>obligatory (bound):</td>
<td>obligatory (bound):</td>
</tr>
<tr>
<td>• Understanding the IIS’ UoD</td>
<td>• Specifying an Information Need</td>
<td></td>
</tr>
<tr>
<td>• Extracting Information from Data Sources</td>
<td>• Understanding Information Sources</td>
<td></td>
</tr>
</tbody>
</table>

**Search Semantics – Matching**
- obligatory (bound): Semantically Exact
- obligatory (bound): Syntactically Approximated
- obligatory (bound): Semantically Approximated

### Selecting a Reference Architecture

- **Selected Features**
- **Matching Algorithm**

**Matching of Feature Nodes**

<table>
<thead>
<tr>
<th>Node-match</th>
<th>Bound</th>
<th>Common</th>
<th>Undecided</th>
<th>Uncommon</th>
<th>Removed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bound</td>
<td>☑ 1</td>
<td>☐ 0,75</td>
<td>☐ 0,5</td>
<td>☑ 0,25</td>
<td>☑ ☑ ☑ 0</td>
</tr>
<tr>
<td>Removed</td>
<td>☑ ☑ 0</td>
<td>☑ ☑ 0,25</td>
<td>☑ ☑ 0,5</td>
<td>☑ ☑ 0,75</td>
<td>☑ ☑ ☑ ☑</td>
</tr>
<tr>
<td>Undecided</td>
<td>☑ ☑ ☑</td>
<td>☑ ☑ ☑</td>
<td>☑ ☑ ☑</td>
<td>☑ ☑ ☑</td>
<td>☑ ☑ ☑</td>
</tr>
</tbody>
</table>

**Similarity of Feature Trees**

\[
\text{match}(f_{F1}, f_{F2}) = \text{node-match}(f_{F1}, f_{F2}) \cdot \Sigma \frac{1}{n} \cdot \text{match}(f_{iF1}, f_{iF2})
\]

- Subfeatures have the same priority
- Details of undecided features are not considered
- Details of removed features are not considered
Design based on Reference Architectures

Outline

• Information Integration Systems (IIS) @ TUB
• How to Support the Design of IIS?
• Feature-Based Design: Reference Architectures
  • Feature-Based Classification of IIS
  • Feature Matching
• Features for ... – Some Ideas
• Conclusions

Features for ...

• Query Processing and Optimization
• Improving Query Results
  - by Improved Schema Mappings
  - by Query Relaxation
  - by Advanced Ranking
  - by Clustering of Results
• Crawling / Indexing

Idea 1: Ontology Languages / Algorithms

RDM / OWL / ...

Algorithm to ...
Idea 1: Ontology Languages / Algorithms

1. Selecting appropriate languages and algorithms
   - Matching of IIS specification and pattern specification
2. Validating manually combined languages and algorithms
   - Matching of pattern specifications

Idea 2: Techniques Improving Query Processing

- Classifying Query Processing Algorithms and Improvements using
  - Statistical Information of Data Sources
  - Additional Semantic Knowledge
- Specifying Relationships to IIS Quality Features

Idea 2a: Metadata Components

- Thesaurus
- Pattern
- Model Trafo
- How is it used?

Idea 2b: Self-Tuning Integration Layer

- Query Optimization
- Requirements
  - Matching
  - Algorithm xy.
- Statistics
  - Queries
  - Amount of Data
  - Response Time
  - ...
Conclusions

Feature Modelling allows to control reuse
- Structures possible features and specifies dependencies
- Allows to characterize existing patterns with respect to functional and quality features
- Allows to define
  - feature matching (what can be reused?)
  - model transformation rules (how is it used?)

Feature-based classification of IIS
- IIS reference architectures
- Query processing and optimization / Improving query results
- Techniques from both schema-based systems and IR-based systems

Current and Future Work

Platform-Independent Design of IIS
- Integration of Data Sources
- Query Processing and Optimization
- Information Integration / Object Fusion

Model-Driven Engineering
- Feature Modelling
- Model Management
- Model Transformation
- Feature Discovery

References (1)


References (2)