Quality-driven Integration of Heterogeneous Information Sources

J.-C. Freytag\textsuperscript{1}
F. Naumann\textsuperscript{1}, U. Leser\textsuperscript{2}

\textsuperscript{1} Humboldt University Berlin
\texttt{(naumann,freytag)@dbis.informatik.hu-berlin.de}
\textsuperscript{2} Technical University Berlin
\texttt{leser@cs.tu-berlin.de}

Motivation

**Query:** Give me all genes on the human X chromosome with pattern `...` in their sequence

- Many alternative ways of obtaining answers
- Varying information quality

⇒ Find plans with highest information quality

Similar problems in:
- Travel planning systems
- Stock information services
- etc.
Framework

Plan: combination of “wrapper” queries

Our solution: 1. Preparation

Annotate correspondences with information quality:
2. Source selection

Use IQ scores to remove the worst sources:

![Diagram of source selection process]

3. Query planning

Compute all plans:

![Diagram of query planning process]

\[ \rho_1 := q_3 q_5 q_7 q_9 \ldots \]
\[ \rho_2 := q_3 q_5 q_7 q_9 \ldots \]
\[ \rho_3 := q_3 q_5 q_7 q_9 \ldots \]
\[ \text{etc.} \]
4. Plan selection

Use IQ scores to find the best plans:

\[
p_1 := q_1 q_2 q_3 q_4 \ldots
\]
\[
p_2 := q_2 q_3 q_4 q_5 \ldots
\]
\[
p_3 := q_3 q_4 q_5 q_6 \ldots
\]

etc.

\[\text{IQ}(p_j) = (60,0.3\ldots) \lor (46,0.2\ldots) \lor \ldots\]

Overview

- Introduction
- Describing Information Sources
- Information Quality: Criteria & Classification
- Finding High-Quality Plans
- Discussion
Query Correspondence Assertions

- Connect two conjunctive queries:
  - one executable query against the aggregate schema of some wrapper
  - one intensionally equivalent query against the mediator schema
  - head is “shared”

Examples

The global schema:

Some exemplary QCAs:

QCA1: sequence(Gn,Se,Or,An) ← S2·v1(Gn,Se,Or,An) ← seq(Gn,Se,Or,An)
QCA2: sequence(Gn,Se,Or,An) ← S2·v1(Gn,Se,Or,An) ← seq(Gn,Se,Or,An)
QCA3: gene(Gn→),cDNACluster(Gn,Dn),cDNA(Dn,Ch,P1,P2) ← S1·v1(Gn,Dn,Ch,P1,P2) ←
   clustering(Gn,Dn,Ch,P1,P2)\_cluster(Ch),primers(Dn,P1,P2)
QCAs allow ...

IQ score dependent refinement:
- $S_2 \cdot v_1$ is "good" for human sequence data
- $S_2 \cdot v_1$ is "bad" for mouse sequence data

Overview

- Introduction
- Describing Information Sources
- Information Quality: Criteria & Classification
- Finding High-Quality Plans
- Discussion
Information quality

- “Fitness for use”
- IQ criteria
  - objective/subjective
  - application domain specific
  - measurement methods
- Empirical study found
  15 criteria [Wang, Strong 96]

IQ criteria can be ...

- Source-specific
  - Understandability
  - Reputation
  - Reliability
  - Timeliness
- QCA-specific
  - Availability
  - Price
  - Rep. consistency
  - Response time
  - Accuracy
  - Relevancy
- Attribute-specific
  - Completeness
  - Amount

Phase 1: Source selection
Phase 2: Query planning
Phase 3: Plan selection
### IQ scores (example)

\[ IQ(QCA) = (Un, Rep, Rel, Ti, Av, Pr, RC, RT, Ac, Re); \]

<table>
<thead>
<tr>
<th></th>
<th>( s_1 )</th>
<th>( s_2 )</th>
<th>( s_3 )</th>
<th>( s_4 )</th>
<th>( s_5 )</th>
<th>( s_6 )</th>
<th>( s_7 )</th>
<th>( s_8 )</th>
<th>( s_9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Underrat.** | 5 | 7 | 7 | 8 | 8 | 6 |       |       |       |
| **Reputation** | 5 | 7 | 7 | 8 | 8 | 7 |       |       |       |
| **Reliability** | 2 | 6 | 4 | 6 | 6 | 6 |       |       |       |
| **Timeliness** | 80 | 80 | 2 | 2 | 1 | 7 |       |       |       |

- **Availability**
  - \( s_1 \): 99
  - \( s_2 \): 99
  - \( s_3 \): 60
  - \( s_4 \): 80
  - \( s_5 \): 95
  - \( s_6 \): 95

- **Price**
  - \( s_1 \): 0
  - \( s_2 \): 0
  - \( s_3 \): 0
  - \( s_4 \): 0
  - \( s_5 \): 0

- **R-Consist**
  - \( s_1 \): 1
  - \( s_2 \): 1
  - \( s_3 \): 0.5
  - \( s_4 \): 0.7
  - \( s_5 \): 0.7

- **Resp-Time**
  - \( s_1 \): 0.2
  - \( s_2 \): 0.2
  - \( s_3 \): 0.2
  - \( s_4 \): 3
  - \( s_5 \): 0.1

- **Accuracy**
  - \( s_1 \): 99.9
  - \( s_2 \): 99.9
  - \( s_3 \): 99.8
  - \( s_4 \): 99.95
  - \( s_5 \): 99.95
  - \( s_6 \): 99.95

- **Relevancy**
  - \( s_1 \): 60
  - \( s_2 \): 80
  - \( s_3 \): 90
  - \( s_4 \): 80
  - \( s_5 \): 80
  - \( s_6 \): 60

---

### Overview

- Introduction
- Describing Information Sources
- Information Quality: Criteria & Classification
- Finding High-Quality Plans
- Discussion
Phase 1: Source selection (sketch)

**Input**
- User query, sources with QCAs, and IQ scores
- Data Envelopment Analysis (DEA)
- Only source specific IQ criteria

Phase 1

Source Selection

Phase 2: Query Planning (sketch)

- “Answering queries using views” [Levy et al. 95]
- User query q
- Search plans p: p ⊆ q

⇒ All correct plans:
  - \( p_1 := QCA_1 \land QCA_2 \land QCA_5 \)
  - \( p_2 := QCA_3 \land QCA_4 \land QCA_6 \land QCA_2 \)
  - ...

Phase 1

Source Selection

Phase 2

Query Planning
Phase 3: Plan selection

- Goal: Find best \( N \) plans
- Three steps for each plan
  1. Determine attribute-specific IQ scores
     \( \Rightarrow \) Complete IQ vector per QCA
  2. Aggregate the IQ scores along the plan
     \( \Rightarrow \) Complete IQ vector per plan
  3. Find overall IQ score
     \( \Rightarrow \) IQ scalar per plan

Output
- Quality-ranked plans
- Phase 3
- Plan selection

Step 1: Determine IQ scores

IQ vector for each QCA in a plan
\[
IQ(QCA_2) = (99, 0, 1, 0.2, 99.9, 80, \ldots, \ldots)
\]

- Amount
- Completeness considers:
  - Completeness of each exported attribute
  - Importance of this attribute in user query
  \( \Rightarrow \) Average weighted QCA - completeness

\[
IQ(QCA_2) = (99, 0, 1, 0.2, 99.9, 80, \ldots, 52.8, 0)
\]
Step 2: Aggregate IQ scores

Merge functions:
- Availability: *
- Price: +
- Response time: max
- Completeness: Sylvester
- ...

Plan P₃

QCA₂

P₃: (89.35, 0, 1, 1, 99.8, 28.8, 76.06, 3)

P₃: (1, 1, 0, 1, 1, 0.62, 0.51, 1)

\[ IQ(P_i) = \sum_{j=1}^{n} w_j v_{ij} \]

\[ IQ(P_3) = 0.77 \]

Step 3: Rank plans

Simple Additive Weighting (SAW)

\[ v_j = \frac{d_j - d_j^{\min}}{d_j^{\max} - d_j^{\min}} \]

\[ v_j = \frac{d_j^{\max} - d_j}{d_j^{\max} - d_j^{\min}} \]
Results

Order of plans depends on:
- IQ scores of sources,
- IQ scores of QCAs and attributes
- User weighting of criteria
- User weighting of attribute importance
- [length]

Order of plans does not depend on:
- Join order

Discussion

- Fine-grained IQ score assignments:
  - per source, per query and per attribute
  - selection and classification of criteria

- Quality as "cost":
  - different levels
  - merge functions

- Outlook:
  - branch & bound algorithm
  - IQ of union operations
Quality-driven Integration of Heterogeneous Information Sources

Questions?