Schema Matching in a Large Scale
Personal Schema Based Querying

Marko Smiljanić, Maurice van Keulen, Willem Jonker

Dutch-Belgian Database Day - December 3, 2004 - Antwerp, Belgium
in this talk

• motivation
  personal schema based querying

• understanding
  formalizing the schema matching problem

• solving
  clustering in schema matching

• validating
  semantic validation without semantics
mediated schema
personal schema

PSQ – Personal Schema Based Query Answering System
Déjà Vu

- personal schema, personal query
- ranked set of mappings
- the answer

Automatic Schema Matching

- Automatic Schema Matching. Semantic mappings are expressions that relate
goals and issues

goals

• efficiency of schema matching  
  (time-to-last, time-to-first)

• effectiveness of schema matching  
  (precision/recall)

issues

• trees vs. graphs

• the objective function
understanding
schema matching

hints
formalism

constraint optimization problem

\[ P = (X, D, C, \Delta) \]

- \( X = (x_1, \ldots, x_n) \) is a list of variables,
- \( D = (D_1, \ldots, D_n) \) is a list of finite domains
- \( C = \{c_1, \ldots, c_k\} \) is a set of constraints, \( c_i : D \rightarrow \{\text{true, false}\} \)
- \( \Delta : D \rightarrow \mathbb{R} \) is the objective function

well known framework, offering a range of approaches for efficient problem solving
$P = (X, D, C, \Delta)$

- $X = (x_1, \ldots, x_n)$
- $D = (D_1, \ldots, D_n)$
- $C = \{c_1, \ldots, c_k\}$
- $\Delta : D \rightarrow \mathbb{R}$
finding a solution
the idea of clustering

distance based clustering
why clustering?

- clusters can be ranked
- search space is reduced
clustering approaches (and issues)

- clustering method has to be scalable

**k-medoid**

- how to initialize
- pre-computation of distance

**hand made linear-time clustering**

- make it intelligent,
  yet keep it close to linear-time
validation paradox

- **semantic validation** does not like large search spaces!

  VS.

- **clustering** is only useful in large search spaces!

\[ P = \frac{T}{A} \]
\[ R = \frac{T}{H} \]
estimating the precision and recall

• size based
• order based
size based quality estimation

no clustering

\[ P = \frac{T}{A} \]

\[ R = \frac{T}{H} \]

yes clustering

\[ R_{12} = \frac{B}{A} \]
size based quality estimation

Precision versus recall curve

NO CLUSTERING

CLUST. BEST CASE

CLUST. WORST CASE

B/A = 93%
order based quality estimation

\[ A_S(x) \]

\[ A_{S_2}(x) \]

no clustering

yes clustering

\[ \mathcal{P}_H_{S_2}(\delta) = \mathcal{P}_H_{S_1}(\delta') \cdot \mathcal{P}_T_{12}(\delta) + \mathcal{P}_H_{S_1}(\delta', \delta) \cdot (1 - \mathcal{P}_T_{12}(\delta)) \]
order based quality estimation

Precision versus recall curve

NO CLUSTERING

CLUST. ALG 1

CLUST. ALG 2
what comes next

• add intelligence to clustering
• impact of other hints on clustering
• using graphs
En dat was het dan!

Vragen?