INCREMENTALLY UNDERSTANDING BIG DATA

Christoph Pinkel
fluid Operations AG / DHBW Mannheim
THE TROUBLE WITH “BIG”
THE TROUBLE WITH “BIG”
THE OTHER “BIG” Problem
MOTIVATING SCENARIO

User (engineer) 

Information Need 

IT Expert 

disparate sources
MOTIVATING SCENARIO

User (engineer) → IT Expert → disparate sources

Translation
MOTIVATING SCENARIO

Query Formulation

Validation

disparate sources

Query Translation
MOTIVATING SCENARIO

Oil drilling use case (Statoil):
- DB 1: 10 TB
- DB 2: 100 TB
- 3,000 Tables
  [DB 3: 2 PB]

disparate sources
MOTIVATING SCENARIO

Query Formulation

Validation

Takes days or **weeks**

Query Translation

disparate sources
MOTIVATING SCENARIO

Query Formulation

Transformation & Execution

Ontology & Mappings

disparate sources
MOTIVATING SCENARIO

Query Formulation

Transformation & Execution

New Schema: **Ontology**

Mapping

Old Schema (Source Schema)

Ontology & Mappings

disparate sources
MOTIVATING SCENARIO

Query Formulation

Transformation & Execution

Ontology & Mappings

disparate sources

Optique
MOTIVATING SCENARIO

Query Formulation

Transformation & Execution

Ontology & Mappings

do disparate sources

Should take only **hours**
MOTIVATING SCENARIO

Ontology & Mappings

How?

disparate sources
MOTIVATING SCENARIO

Ontology & Mappings

disparate sources
MOTIVATING SCENARIO

Ontology & Mappings

Incremental Approach!

disparate sources
AGENDA

• Motivation

• **Schema Mapping – State of the Art**
  • Semi-Automated Incremental Approach

• Outlook & Conclusion
THE MAPPING PROBLEM
THE MAPPING PROBLEM

New

Old
MANUAL MAPPINGS

Old

New

?
MANUAL MAPPINGS

New

Old
MANUAL MAPPINGS

New

Old
PAY AS YOU GO

“Know exactly what you need”

“Do precisely what you must”
PARTIAL MANUAL MAPPINGS

New

Old
PARTIAL MANUAL MAPPINGS

If required new schema precisely known!
SEMI-AUTOMATIC MAPPINGS
SEMI-AUTOMATIC MAPPINGS

New

Old

0.3 0.9 0.7 0.5 0.4 0.6 0.8 0.3 0.5

0.9 0.5 0.4 0.6 0.9 0.3 0.5
SEMI-AUTOMATIC MAPPINGS

New

Old

0.3

0.9

0.7

0.5

0.4

0.6

0.8

0.3

0.5
SEMI-AUTOMATIC MAPPINGS

New

Old

0.3

0.9

0.7

0.5

0.4

0.6

0.8

0.9

0.3

0.5
SEMI-AUTOMATIC MAPPINGS

New

Old

0.3

0.9

0.7

0.5

0.4

0.6

0.8

0.3

0.5
SEMI-AUTOMATIC MAPPINGS

New

Old

0.3 0.9 0.7 0.5 0.4 0.6 0.8 0.3 0.5

✔ ✘
SEMI-AUTOMATIC MAPPINGS

New

Old

0.3

0.9

0.7

0.6

0.8

0.4

0.5

0.9

0.3

0.5
CROSS PRODUCT TROUBLE
CROSS PRODUCT TROUBLE

New

Old

disparate sources
SEMI-AUTOMATIC PARTIAL MAPPINGS

- Reduce size of cross product
- Still need precise definition of target schema

- Reduce number of manual choices
- Still need to pick/confirm for each part separately
SEMI-AUTOMATIC PARTIAL MAPPINGS

- Reduce size of cross product
- Still need precise definition of target schema

- Reduce number of manual choices
- Still need to pick/confirm for each part separately
SEMI-AUTOMATIC, GRAPH SUPPORTED (MELNIK ET AL.)
SEMI-AUTOMATIC, GRAPH SUPPORTED (MELNIK ET AL.)
SEMI-AUTOMATIC, GRAPH SUPPORTED (MELNIK ET AL.)

New

Old

0.9
SEMI-AUTOMATIC, GRAPH SUPPORTED (MELNIK ET AL.)
SEMI-AUTOMATIC, GRAPH SUPPORTED (MELNIK ET AL.)
SEMI-AUTOMATIC, GRAPH SUPPORTED (MELNIK ET AL.)

New

Old

0.9

0.8

0.3

0.5

0.4

0.6

0.8

0.3

0.9

0.2
AGENDA

• Motivation
• Schema Mapping – State of the Art
• **Semi-Automated Incremental Approach**
• Outlook & Conclusion
VISION

Incremental Understanding and Mapping of Data

Pay As You Go
Query Driven
Semantics in Ontology
Exploitation of Partial Mappings

Schema Extraction
OBDA
Semi-Automatic Mappings
Pay As You Go
Query Driven
Incremental Understanding and Mapping of Data
Semantics in Ontology
Exploitation of Partial Mappings
Schema Extraction
OBDA
Semi-Automatic Mappings
BUILD ON EXISTING APPROACHES

Old

New

0.3

0.9

0.5

0.7

0.4

0.6

0.8

0.3

0.5

0.9
BUILD ON EXISTING APPROACHES
BUILD ON EXISTING APPROACHES
BUILD ON EXISTING APPROACHES

Old

New

0.3

0.0

1.0

0.5

0.4

0.6

0.8

0.3

0.5

0.9
BUILD ON EXISTING APPROACHES
INCREMENTAL PROCESS

Information Need

IncMap
INCREMENTAL PROCESS

Mapping Suggestions

IncMap
INCREMENTAL PROCESS

Feedback
INCREMENTAL PROCESS

IncMap

Results
INCREMENTAL PROCESS

IncMap

Partial Mapping

Mapping Meta Data
INCREMENTAL PROCESS
INCREMENTAL PROCESS
EMPLOYED USER FEEDBACK

• Confirmation/rejection of mappings
  • Improve initial scores in similarity graph
  • Can be re-used, possibly weighted by user confidence

• Confirmation/rejection of sample results
  • Gives indications of several likely mappings
  • Increases/decreases likelihood of related mappings
EMPLOYED USER FEEDBACK

• Confirmation/rejection of mappings
  • Improve initial scores in similarity graph
  • Can be re-used, possibly weighted by user confidence

• Confirmation/rejection of sample results
  • Gives indications of several likely mappings
  • Increases/decreases likelihood of related mappings
MODIFIED MATCHING ALGORITHM

Melnik

IncMap

1.0
MODIFIED MATCHING ALGORITHM

Melnik
0.5

IncMap
1.0

0.6
MODIFIED MATCHING ALGORITHM

Melnik

IncMap

0.5

1.0

0.6
ADVANTAGES

• Little initial effort

• Even for full eventual matches of schema:
  • Improved results through intermediate user feedback
  • Easier to update/extend

(sum of parts take less effort than creating all at once)
AGENDA

• Motivation

• Schema Mapping – State of the Art

• Semi-Automated Incremental Approach

• Outlook & Conclusion
OUTLOOK

• Prototype needs to prove use in practice

• Need extensions for more flexible mappings (1:n / n:1)

• Additional user feedback could be incorporated

➡️ Could eventually lead to principled step forward in incremental understanding of complex data
CONCLUSION

• Severe problems currently exist to formulate information needs on data with complicated structure

• Semantic technologies offer a chance to ease these problems

• Integration/mapping to ontologies requires major effort

• Incremental approach seems appealing

• Experimental results still pending
CONCLUSION

• Severe problems currently exist to formulate information needs on data with complicated structure

• Semantic technologies offer a chance to ease these problems

• Integration/mapping to ontologies requires major effort

• Incremental approach seems appealing

• Experimental results still pending

Questions?
Shoot!