Scalable Matching of Industry Models – A Case Study

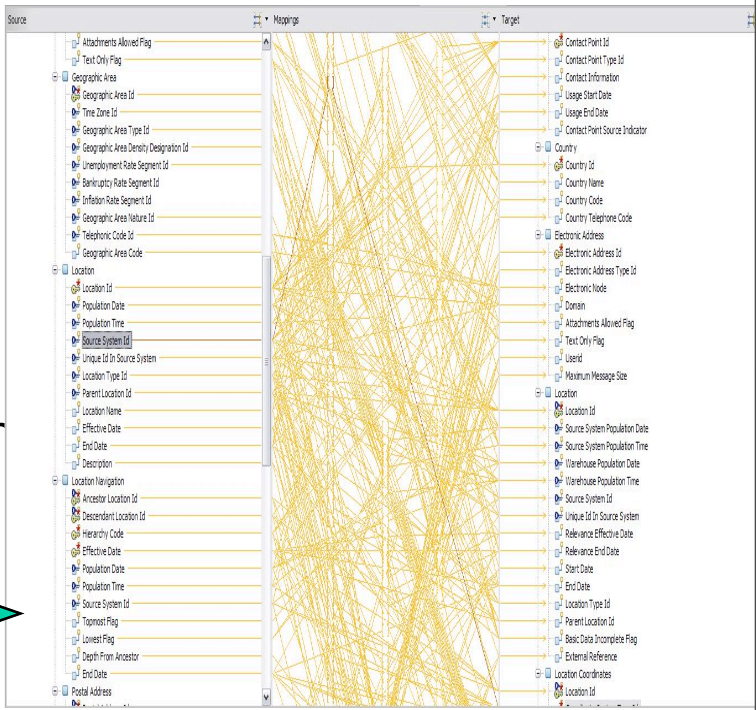
Brian Byrne, Achille Fokoue, Aditya Kalyanpur, Kavitha Srinivas, and Min Wang

© 2009 IBM Corporation

Background and motivation

Problem

- Industry models:
  - Diverse formats (UML, ER, XSD, etc)
  - Multiple aspects: data, processes, services
  - Multiple domains (Healthcare, finance, insurance, etc)
  - Large models
  - Little to no formal semantics
  - Informal semantics buried in documentation (PDF, Excel, etc)
- Existing tools do not scale well to large models
- Reviewing matching is as tedious as developing them.



Result

- Labor intensive matching in solution building
- Poor quality of manual mappings
- No scalable tools for reviewing the quality of mappings.

© 2009 IBM Corporation

Technical Approach

Service50->Sex (in A)

Associated Documentation:
The patient's gender

UML | ER | EMF

Person->Gender (in HPDM)

Associated Documentation:
The Gender of the Person

Fields	Values
Name	Sex
Class	Service50
Documentation	Gender
Documentation	Patient

Virtual Document Builder

Fields	Values
Name	Gender
Class	Person
Documentation	Gender
Documentation	Person

Terms	Freqs
Gender	1
Patient	1
Service50	1
Sex	1

Model Element Similarity Computation

Terms	Freqs
Gender	2
Person	2

Rank 1: iservice50.sex, person.gender --- 0.43
Rank 2: iservice50.sex, body element.gender --- 0.36

Manual mappings → Lint Engine → Suspicious mappings to filter

iservice50.sex, body element.gender

© 2009 IBM Corporation

Integrating lexical and semantic similarity between terms

Terms	TF-IDF
Gender	1
Sex	0

Terms	TF-IDF
Gender	0
Sex	1

© 2009 IBM Corporation

Integrating lexical and semantic similarity between terms

Terms	TF-IDF
Gender	1
Sex	0

Terms	TF-IDF
Gender	0
Sex	1

Cosine = 0

© 2009 IBM Corporation

Integrating lexical and semantic similarity between terms

Terms	TF-IDF
Gender	1
Sex	0

Terms	TF-IDF
Gender	0
Sex	1

Cosine = 0

© 2009 IBM Corporation

Integrating lexical and semantic similarity between terms

Terms	TF-IDF
Gender	1
Sex	0

Terms	TF-IDF
Gender	0
Sex	1

Cosine = 0

Terms	TF-IDF
Gender	1 + 0 * sim(Gender,Sex)
Sex	0 + 1 * sim(Sex, Gender)

Terms	TF-IDF
Gender	0 + 1 * sim(Gender,Sex)
Sex	1 + 0 * sim(Sex, Gender)

© 2009 IBM Corporation

Integrating lexical and semantic similarity between terms

Terms	TF-IDF
Gender	1
Sex	0

Terms	TF-IDF
Gender	0
Sex	1

Cosine = 0

Terms	TF-IDF
Gender	1 + 0 * sim(Gender,Sex)
Sex	0 + 1 * sim(Sex, Gender)

Terms	TF-IDF
Gender	0 + 1 * sim(Gender,Sex)
Sex	1 + 0 * sim(Sex, Gender)

$$\text{Cosine} = \frac{2 * \text{sim(Gender, Sex)}}{1+(\text{sim(Gender,Sex)})^2}$$

© 2009 IBM Corporation

Experiments

- Models tested
  - A (customer model) vs. HPDM model (healthcare)
  - B (customer model) vs. BDW model (finance services)
  - MDM physical model (master data) vs. HPDM model.
  - MDM physical model vs. BDW model.
  - C (customer model) vs. BDW model.
  - RDWM (retail model) vs. BDW model.
  - BDW vs. IAA model (insurance).

These model mappings are frequently requested by customers

- Model selection based on availability of
  - Manually constructed mappings
  - Available domain expert for evaluation of mappings

© 2009 IBM Corporation

Overview of Results

Models	Total # matches	Precision of the top 100
A ->HPDM	43	67%
B ->BDW	197	74%*
MDM->BDW	149	71%*
MDM->HPDM	324	54%*
RDWM->BDW	3632	100%*
C->BDW	3263	96%*
IAA->BDW	69	52%*

*Estimates were based on validation by domain experts because of problems in the quality of manually constructed mappings for 3 of 4 models. Estimates were based on the top 100 mappings.

© 2009 IBM Corporation

Lint Engine: An approach to Improving the Quality of Manual Mappings

- Manual mappings are surprisingly bad for 3/4 models:
  - Contains elements that do not match elements in either model
    - Poor transcription of names (changes of spaces, appending package names, etc).
    - Mapper created new classes/attributes to make up a mapping (e.g., DUMMY.DUMMY_ATTR in AMEX).
  - Contains mappings to an "absurdly" generic class
  - Contains mappings that are just wrong

```
location.location id || zip code territory manager.postal code
condition.condition id || midw fee arrangment.effective date
location.location id || merchant contact.telephone extension
number
condition.condition id || edw discount rate.account transaction
rate
```

© 2009 IBM Corporation

“Lint” for model mapping

Identify heuristics to detect suspicious mappings

- Can be used as a tool to ‘review’ model mappings created by a human
- Can be used on output of the mapping tool to identify groups of suspicious mappings

Example heuristics implemented

- A model element with an exact lexical match was not returned.
- A single element of one model was mapped to multiple elements of another models
- 6 categories implemented

© 2009 IBM Corporation

Lint applied to B-BDW manual mappings

Total number of mappings:	306
Total number of suspicious mappings:	151 (51 %)
Exact Name Not Match:	13 (8 %)
One To Many Mappings:	143 (46 %)
Mapping Without Documentation:	40 (25 %)
Duplicate Documentation:	2 (1%)

© 2009 IBM Corporation

Conclusion and future directions

- Concrete approach to scale model matching to large industry models
- Next steps:
  - Embed semi-automated mapping algorithm into a tool to “suggest” mappings.
  - Incorporate user feedback to teach the algorithm to self correct
    - Utilize machine learning techniques to find the correct ‘features’ for a given model comparison).
    - Reduce variability in automated mapping using “Lint” and machine learning techniques.

© 2009 IBM Corporation

Thanks!

© 2009 IBM Corporation