Searching for a golden needle in association rule mining haystack

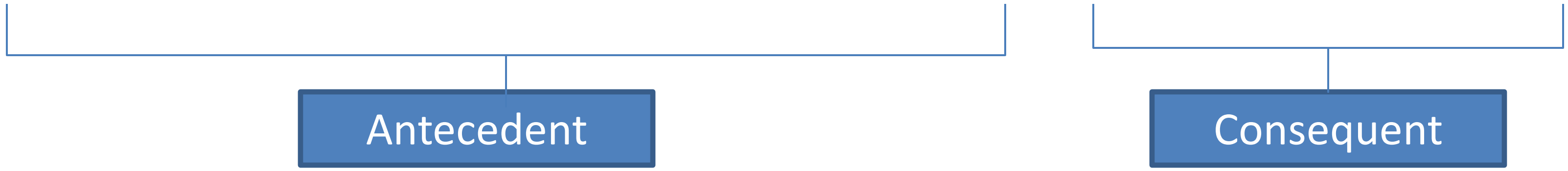
Tomáš Kliegr, Andrej Hazucha, Lukáš Beránek, Marek Ovečka, Tomáš Marek
Department of Information and Knowledge Engineering
Faculty of Informatics and Statistics
University of Economics, Prague

Association Rule Mining

EXAMPLE

Unlike clustering and classification, association rules provide true “nuggets” – rules meeting selected *interest measures*

`Duration(2y+)and District(Prague)=> Loan Quality(good)`



Antecedent

Consequent

THE PROBLEM WITH INTEREST MEASURES

It is usually not possible to tweak the interest measure thresholds so that only the really interesting rules (unknown to the expert) are output. To be on the safe side, we often get (many!) more rules than desired,

SOLUTIONS ATTEMPTED IN SEWEBAR

- **Learn what the expert knows and automatically filter mining results**
  Already published at ISMIS’09: Semantic Analytical Reports, Springer 2009,
- **Get more rules and let the expert search within the rules**
  Presented at this seminar
- **Learn what the expert knows and use this knowledge to narrow the data mining task**
  Ever-Miner Project

Knowledge representation

- The output of mining algorithms is in XML
- Industry standard PMML format
- The **natural way** to search is treat PMML as XML and search it by Xquery
- The **fast** and **lightweight** way is to search it as fulltext
- The most **powerful** way is to semantize the PMML and search it by tolog

Searching association rules with tolog

Tomáš Kliegr
Department of Information and Knowledge Engineering
Faculty of Informatics and Statistics
University of Economics, Prague

PMML is “just” an XML Schema

- Developed for deploying mining models
- Good for migration from one data mining environment to another

But:

- No explicit links between nodes
- Verbose
- Self-contained. Lacks support for
  - Interlinking multiple PMML documents
  - Interlinking PMML with other information

Association Rule Mining Ontology

The ontology is a „semantization“ of PMML XML Schema

DESIGN GUIDELINES

The key design principle was to allow easy transformation of data from PMML to AROn

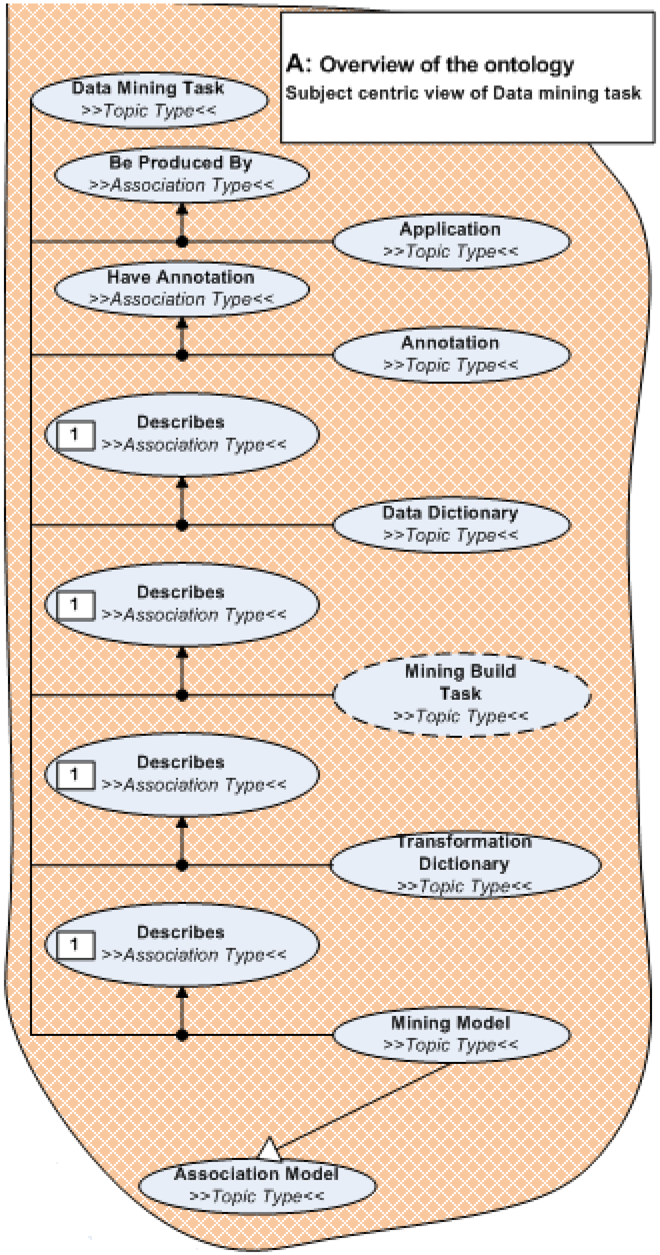
SCOPE

The ontology is limited to the subset of PMML relevant to association rule mining.

60 topic types, 50 association types and 20 occurence types

USE

No automatic transformation is yet available, but we are working on one using OKS framework. Currently, data can be input using Ontopoly.



A: Overview of the ontology
Subject centric view of Data mining task
Data Mining Task >>Topic Type<<
Be Produced By >>Association Type<<
Application >>Topic Type<<
Have Annotation >>Association Type<<
Annotation >>Topic Type<<
1
Describes >>Association Type<<
Data Dictionary >>Topic Type<<
1
Describes >>Association Type<<
Mining Build Task >>Topic Type<<
1
Describes >>Association Type<<
Transformation Dictionary >>Topic Type<<
1
Describes >>Association Type<<
Mining Model >>Topic Type<<
Association Model >>Topic Type<<

The anatomy of an association rule in the Association Rule Mining Ontology

Association Rule

Duration(2y+) and District(Prague)=> Loan Quality(good, medium)

Derived Boolean Attributes

{Duration(2y+) and District(Prague)}

Basic Boolean Attributes

{Duration(2y+); Loan Quality(good, Bad) ; District(Prague)}

Boolean Attributes

{Duration(2y+) and District(Prague) ; Duration(2y+) ; Loan Quality(good, Bad) ;..}

Coefficients

{(2y+); (good), (medium ; Prague)}

Categories

2y+; good; medium; Prague

Basic Boolean Attribute can refer to categories of a Derived Field or a Data Field

Derived Field

Duration, District, Loan Quality

Discretization Bin

Duration(2y+) = duration<24;60>

Value Mapping Bin

Loan Quality(Good) = status(A, B)

Data Field

duration, district, status

Value

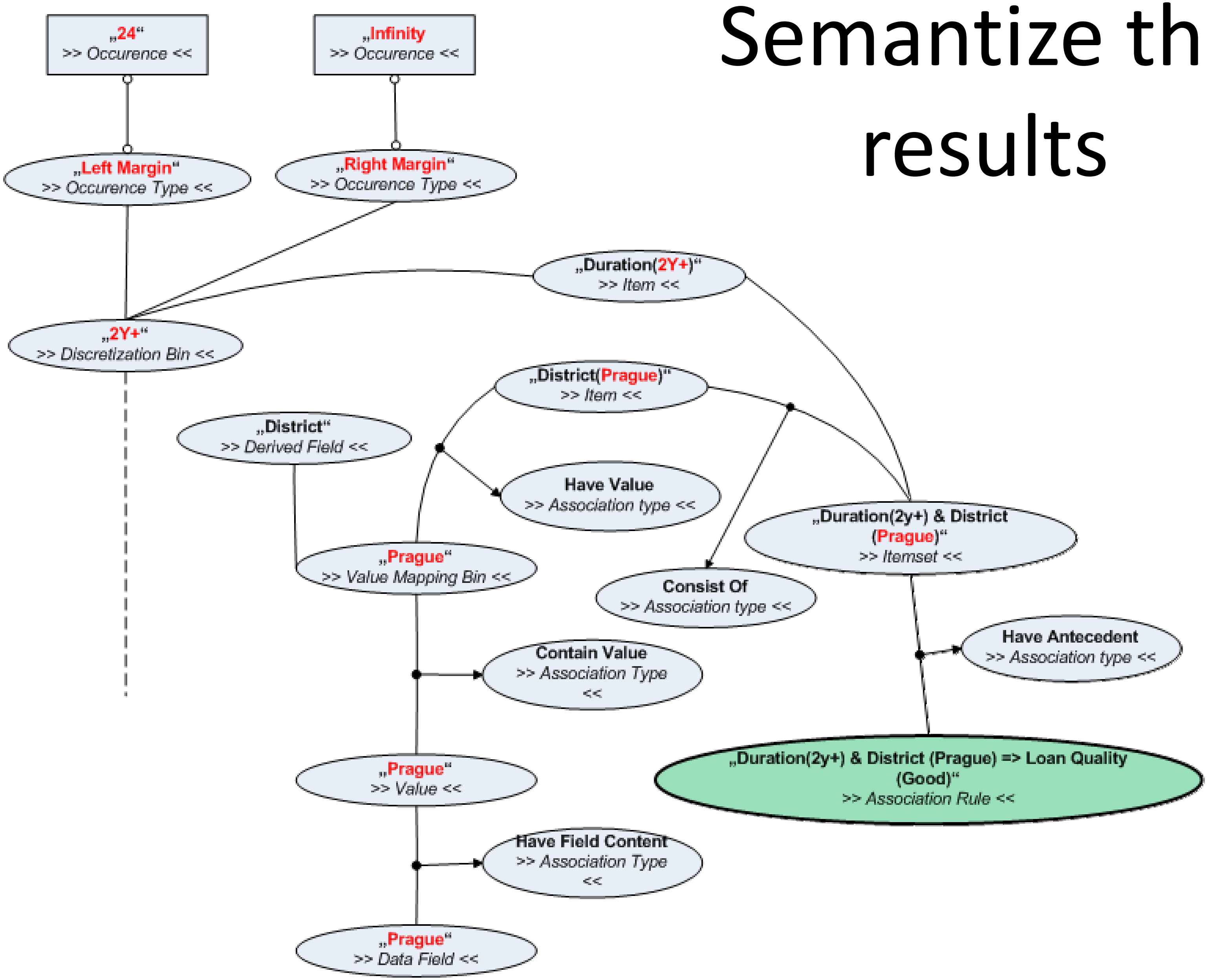
status = {A,B,C,D,E,F}

Interval

duration = <1;60>

$$\text{AR 3: } duration(2y+)\&district(Prague) \Rightarrow statusAggregated(good, medium)$$

Semantize the results

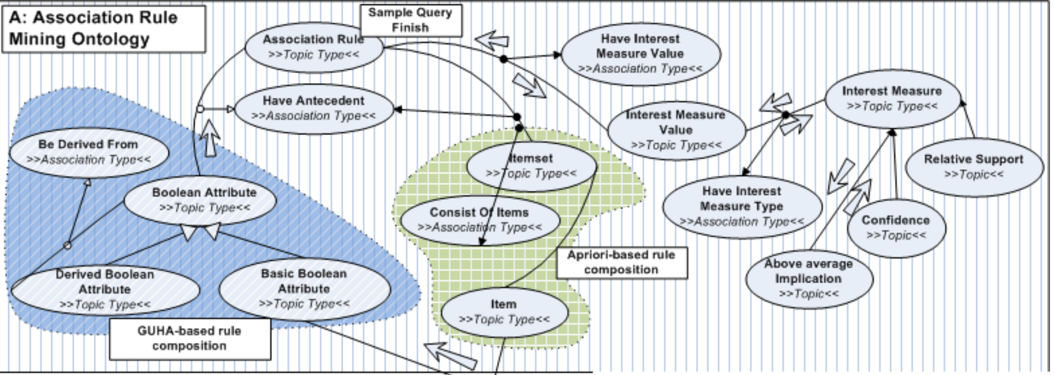


tolog

- Tolog is a topic map query language developed by Ontopia
- Topic maps are interchanged in XML, but their native storage format is often RDBMS
- Syntactically, it is a crossbreed of Prolog and SQL
- Declarative language based on first-order logic

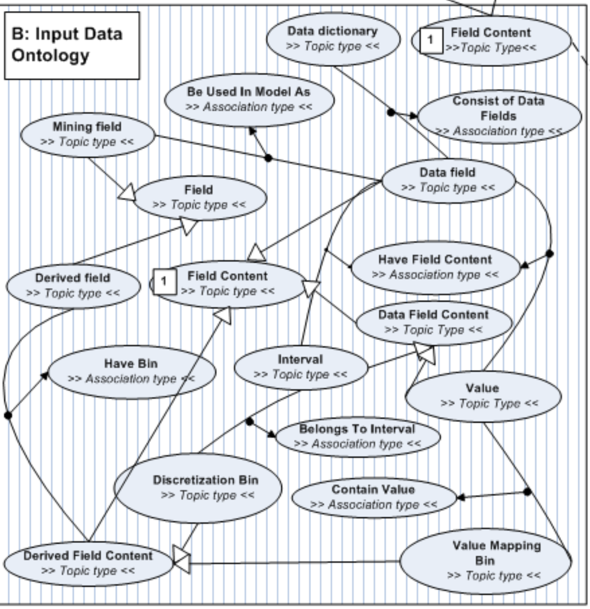
tolog query by example

- Get all association rules containing basic boolean attribute LDL with value “vysoky” in its coefficient in the antecedent
- LDL may refer either to DataField or DerivedField
- The rule must have associated the *Above Average Implication* interest measure


A: Association Rule Mining Ontology
Sample Query Finish
Association Rule >>Topic Type<<
Have Interest Measure Value >>Association Type<<
Have Antecedent >>Association Type<<
Interest Measure Value >>Topic Type<<
Interest Measure >>Topic Type<<
Be Derived From >>Association Type<<
Itemset >>Topic Type<<
Boolean Attribute >>Topic Type<<
Consist Of Items >>Association Type<<
Have Interest Measure Type >>Association Type<<
Relative Support >>Topic<<
Confidence >>Topic<<
Apriori-based rule composition
Derived Boolean Attribute >>Topic Type<<
Basic Boolean Attribute >>Topic Type<<
Above average Implication >>Topic<<
Item >>Topic Type<<
GUHA-based rule composition

Map topics in the query to strings:

```
getTopicForString("vysoky",d:fieldcontent,
    $Cat)
getTopicForString("LDL",d:field, $Field)
getTopicForString("Above Average
Implication",d:4ftquantifier, $Quantifier1)
```


B: Input Data Ontology
Data dictionary >> Topic type <<
1
Field Content >>Topic Type<<
Be Used In Model As >> Association type <<
Consist of Data Fields >> Association type <<
Mining field >> Topic type <<
Data field >> Topic type <<
Field >> Topic type <<
Have Field Content >> Association type <<
Derived field >> Topic type <<
1
Field Content >> Topic type <<
Data Field Content >> Topic Type <<
Interval >> Topic type <<
Have Bin >> Association type <<
Value >> Topic Type <<
Belongs To Interval >> Association type <<
Discretization Bin >> Topic type <<
Contain Value >> Association type <<
Value Mapping Bin >> Topic type <<
Derived Field Content >> Topic type <<

Make sure that vysoky in $Cat belongs to LDL in $Field

A] LDL is a Data Field – or - B] LDL is a Derived Field

```
{d:havefieldcontent($Cat : d:datafieldcontent, $Field:p:DataField)
    |d:havebin($Cat:d:derivedfieldcontent,$Field:p:DerivedField)}
```

Find Basic Boolean Attribute, which has $Cat in its coefficient

```
d:havecoefficient($BBA:d:basicbooleanattribute, $Cat:
    d:coefficient)
```

Get Antecedent, which is derived from BBA

```
GetRuleFromBAinAntecedent($BBA, $RULE)
```

Uses the inference rule:

Either BBA is the antecedent of the rule, or there is a (chain of) boolean attributes derived from it, which is the antecedent of the rule.

```
GetRuleFromBAinAntecedent($BA, $RULE):-
{d:haveantecedent($BA : p:antecedent, $RULE: p:associationrule)|
d:bederivedfrom($DBA:d:derivedbooleanattribute,$BA:d:booleanattribute),
GetRuleFromAntecedentBA($DBA, $RULE)}.
```

Check that the association rule has AA Implication quantifier

```
d:haveinterestmeasurevalue($RULE:p:associationrule,$Val: d:interestmeasurevalue),
d:haveinterestmeasuretype($Val:d:interestmeasurevalue, $Quantifier1 : d:interestmeasure)
```

Conclusion

- Tolog query can be stored in a server side module
- and executed standalone

```
import “arlib.tl” as arlib
[1] arlib:SearchConsequent(“vysoky”, “LDL”, “Above Average Implication”, $Rule)?
```

- or within another query

```
 {arlib:SearchConsequent(“vysoky”, “LDL”, “Above Average Implication”, $Rule) |
    arlib:SearchAntecedent(“vysoky”, “LDL”, “Above Average Implication”, $Rule)}?
```

- or used to construct a new module predicate

```
 arlib:Search (“vysoky”, “LDL”, “Above Average Implication”, $Rule)?
```

- Easy to incorporate background knowledge to the query
- Need to do experiments on larger data
- Several small bugs and missing features in Ontopia (range queries, some queries crash)