Tool-Supported Approaches to Ontology Engineering

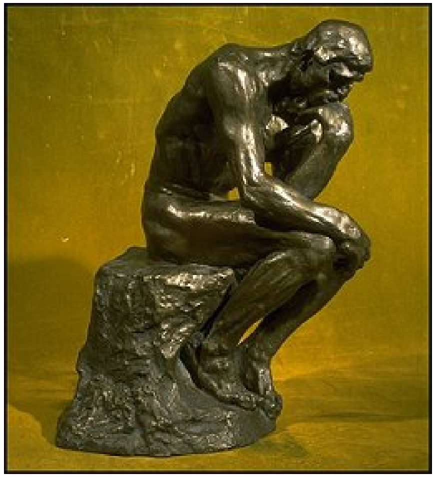
Jon Atle Gulla
Professor of Information Systems
Norwegian University of Science and Technology, Trondheim

Can we generate ontologies automatically?

Some examples of generated ontology structures

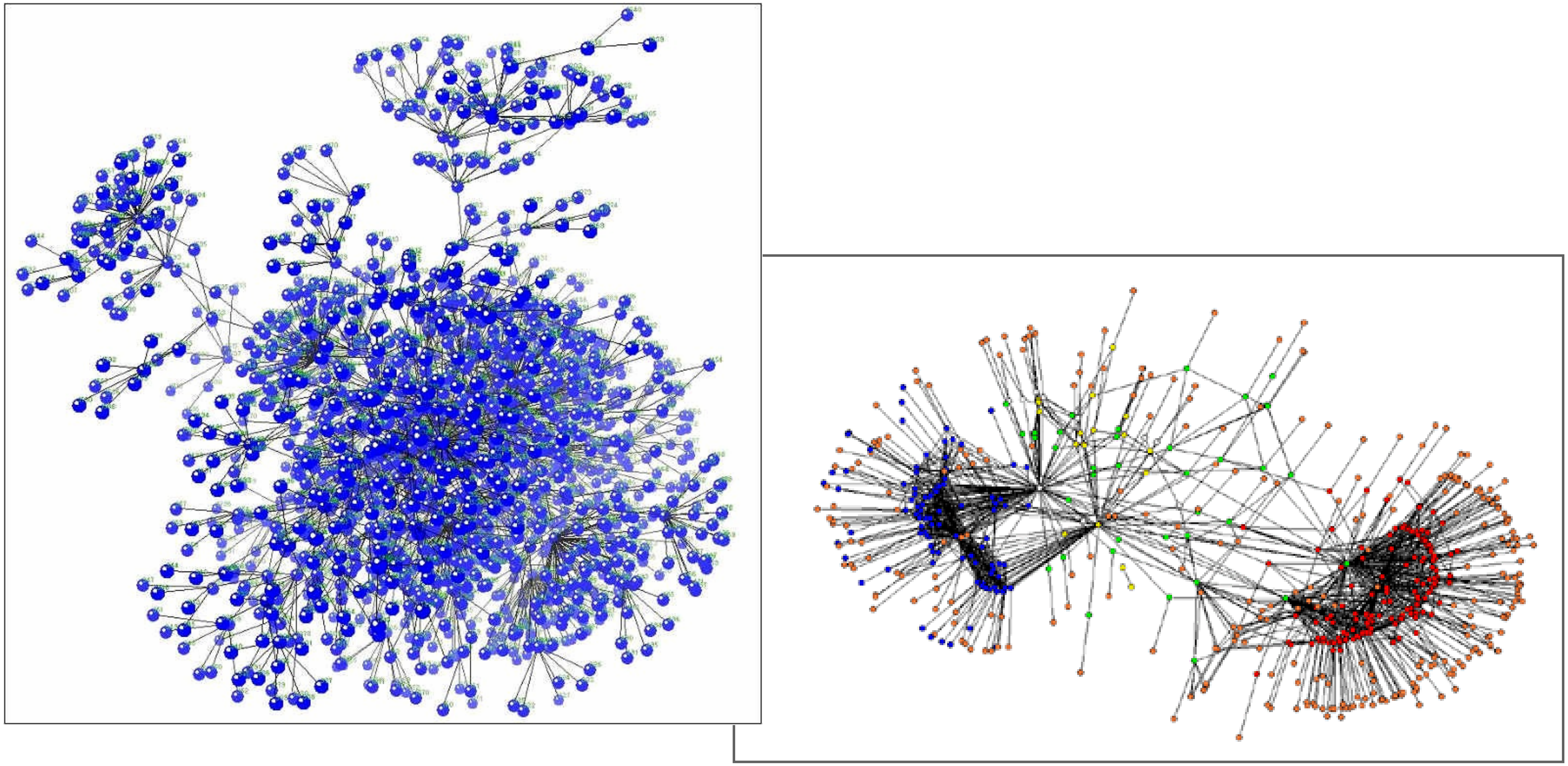
Outline

- Ontology Learning vs. Ontology engineering
- Principles of ontology learning
- Ontology learning strategies
- Learning Classes, Individuals and Relationships in the movie domain
- Quality of ontology learning
- Conclusions



Ontology Engineering

- How to develop and maintain large complex ontologies?



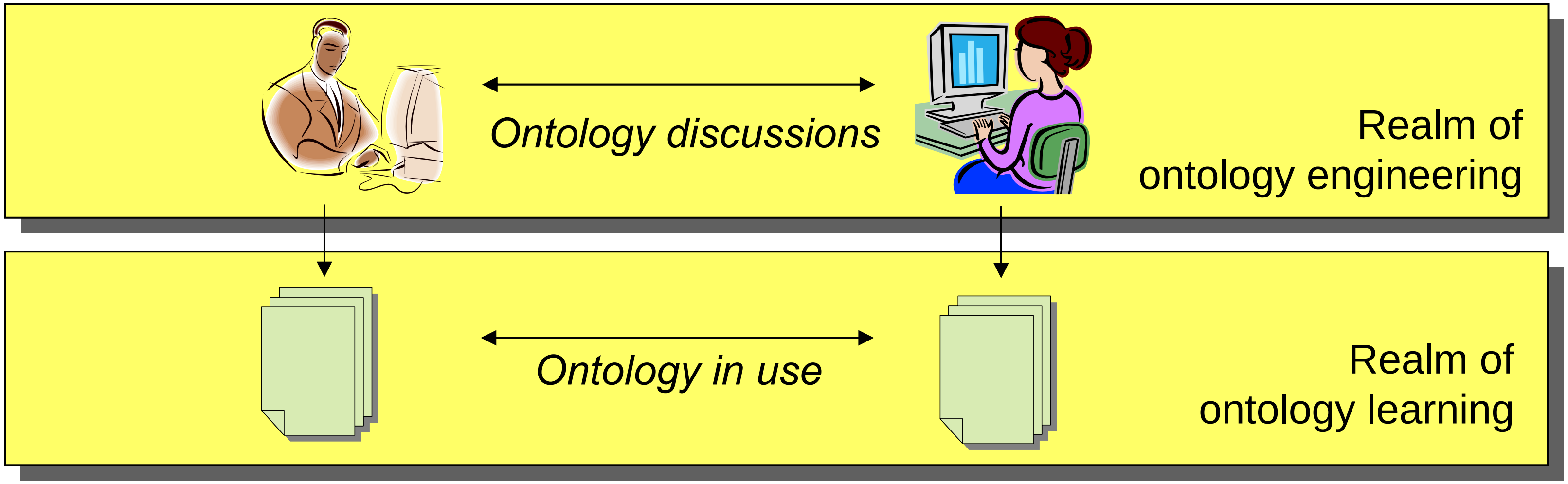
Ontology Modeling vs. Learning

- Traditional ontology engineering approach
  - *Project:* Form team of ontology and domain experts
  - *Ontology & domain experts:* Collaborative manual modeling process
  - *Domain experts:* Verify ontology against domain knowledge
  - *Ontology experts*: Verify ontology against syntactic and semantic quality measures
- Expensive and time-consuming approach

- Ontology learning approach:
  - *Domain experts:* Find representative domain text
  - *Tool:* Extract candidate classes, individuals and properties automatically from domain texts
  - *Ontology & domain experts:* Verify candidate structures and complete ontology
- Can also be used to verify domain quality of existing ontology
- Cost-effective approach

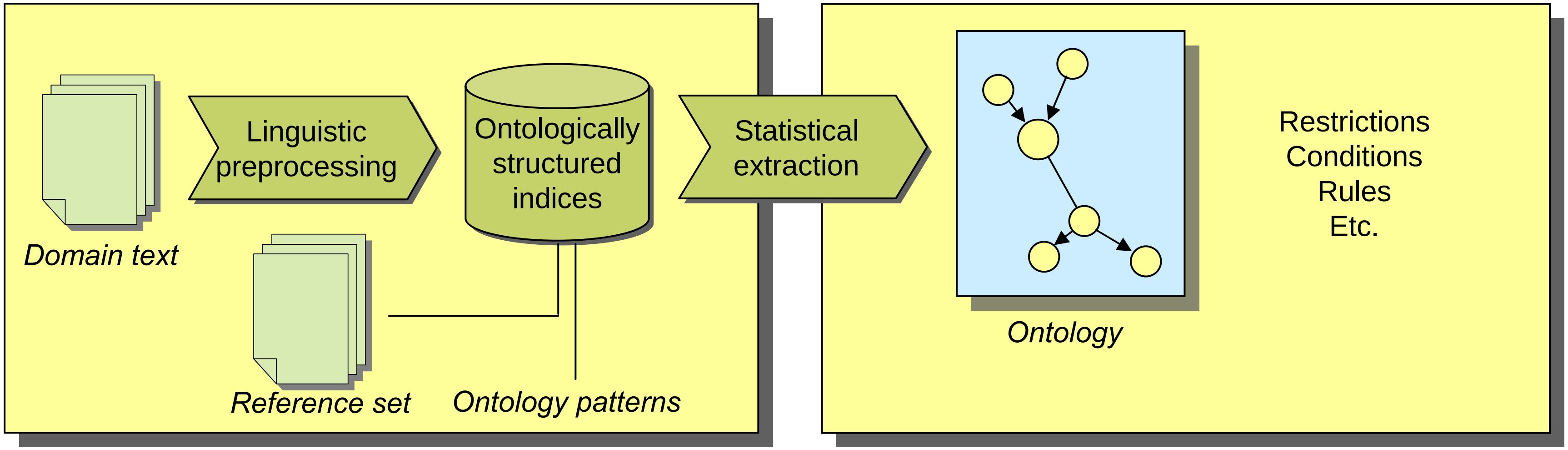
Ontology Learning Basis

- People communicate using domain-specific concepts
- People document using domain-specific concepts
- Ontology learning: *Extract ontology structures from written documentation*



- Requirements:
  - Documents representative for domain terminology
  - Documents cover all the terminology
  - Well-defined and consistent use of terminology in domain

Ontology Learning Process



Domain text

Linguistic preprocessing

Ontologically structured indices

Reference set

Ontology patterns

Statistical extraction

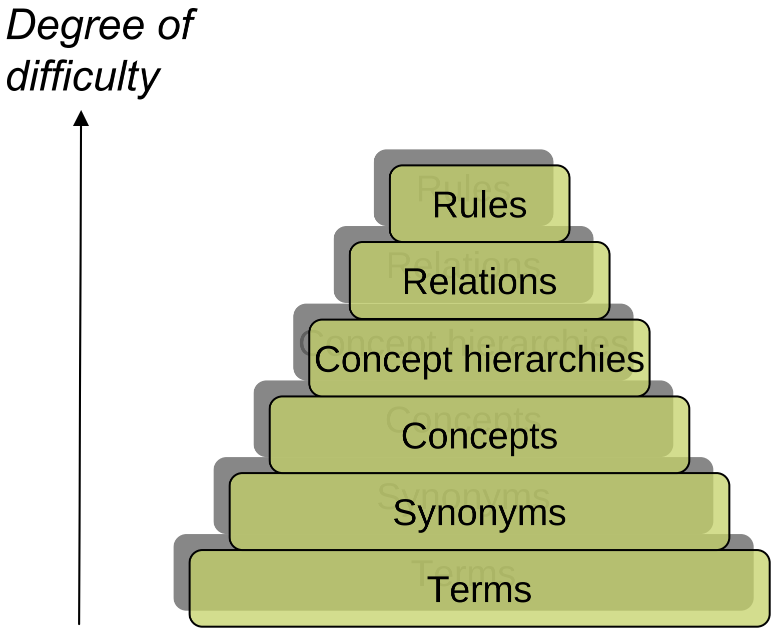
Ontology

Restrictions
Conditions
Rules
Etc.

Automatic extraction of ontology candidate structures

Manual verification of candidates and completion of ontology

Levels of Ontology Learning



$\forall x,y(\mathit{manager}(x,y) \rightarrow \mathit{report}(y,x))$

FINANCE(ag:SPONSOR, go: PROJECT)

is_a(MANAGER, EMPLOYEE)

PROJECT

(leader, manager, lead)

sponsors, costs, charter

Ontology Learning Strategies

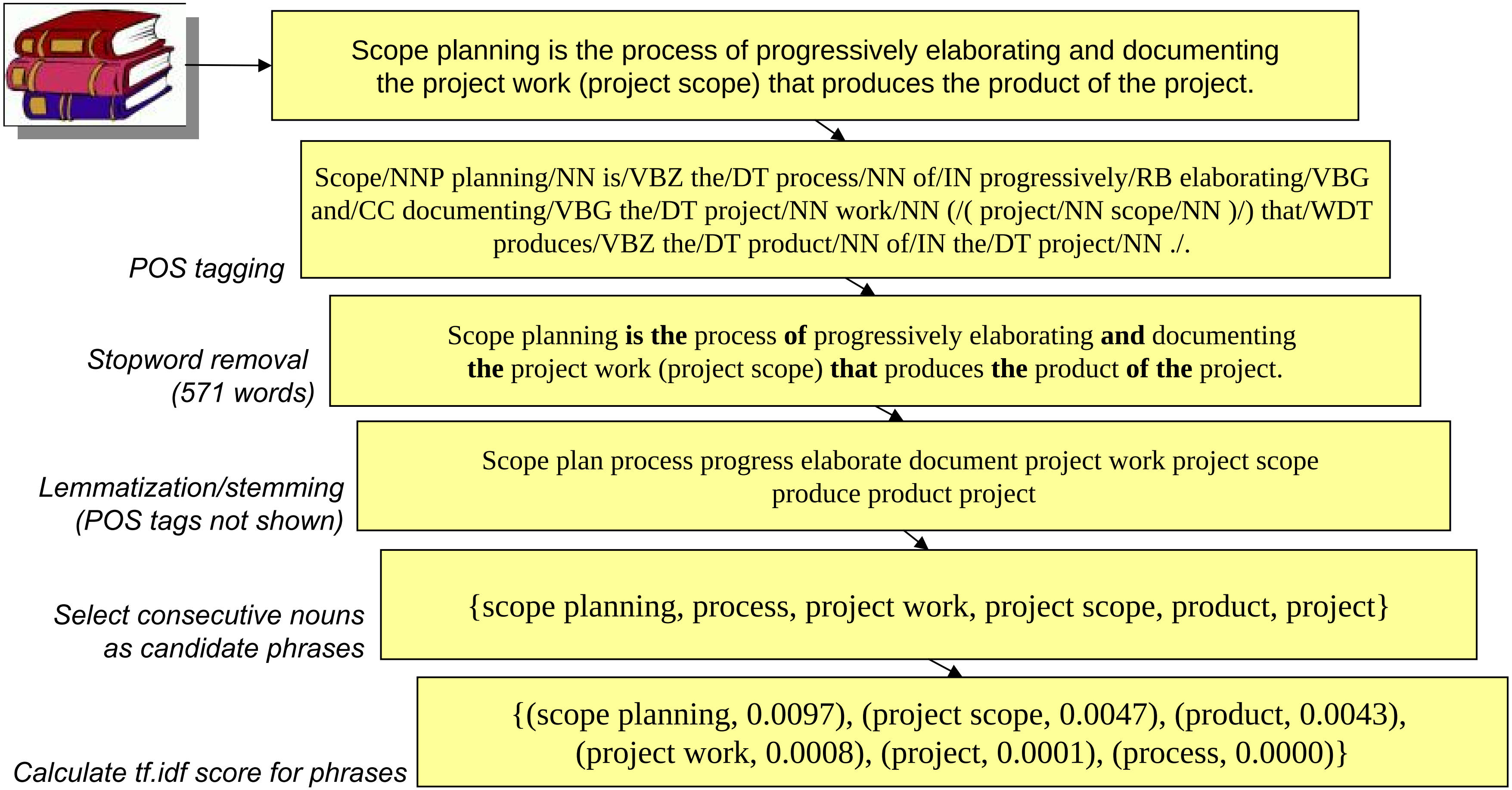
- Term extraction
  - *Linguistic analysis*
  - *Statistical analysis*
- Synonyms
  - *Classification-based techniques*
  - *Distribution-based techniques*
- Concept formation
  - *Structure recognition*
  - *Keyphrase generation*
  - *Instance learning*
- Concept hierarchy
  - *Clustering*
  - *Lexico-syntactic patterns*
  - *Head-modifier approaches*
  - *Subsumption approaches*
  - *Classification-based techniques*
- Relations
  - *Association rules*
  - *Concept vectors*
- Rules
  - *Structure recognition for meta-property recognition*
  - *Dependency trees and path similarities*

Examples: Learning Classes, Individuals and Relationships

Core techniques for ontology learning

Domain: Movie industry
Web data sources: IMDB, Videoload, Wikipedia, etc.
Resulting ontology: Semantic search application

Keyphrase Extraction for Learning Classes



Scope planning is the process of progressively elaborating and documenting the project work (project scope) that produces the product of the project.

POS tagging

Scope/NNP planning/NN is/VBZ the/DT process/NN of/IN progressively/RB elaborating/VBG and/CC documenting/VBG the/DT project/NN work/NN (/(project/NN scope/NN)/) that/WDT produces/VBZ the/DT product/NN of/IN the/DT project/NN ./.

Stopword removal (571 words)

Scope planning **is the** process **of** progressively elaborating **and** documenting **the** project work (project scope) **that** produces **the** product **of the** project.

Lemmatization/stemming (POS tags not shown)

Scope plan process progress elaborate document project work project scope produce product project

Select consecutive nouns as candidate phrases

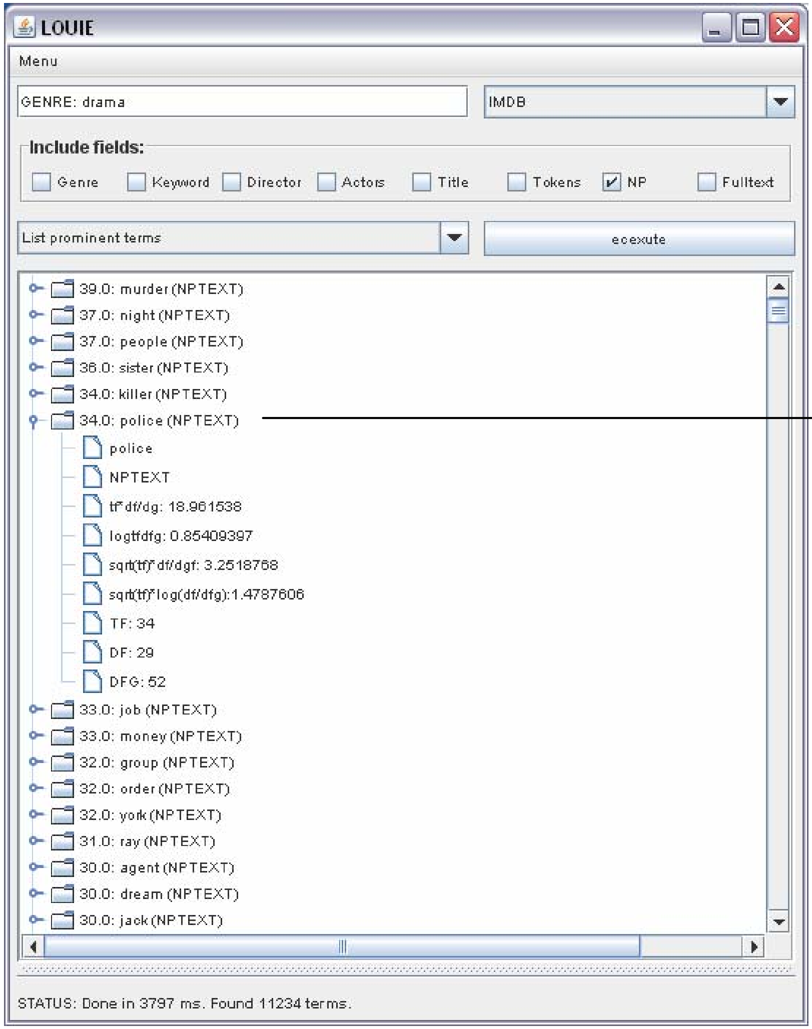
{scope planning, process, project work, project scope, product, project}

Calculate tf.idf score for phrases

{(scope planning, 0.0097), (project scope, 0.0047), (product, 0.0043), (project work, 0.0008), (project, 0.0001), (process, 0.0000)}

$$\mathrm{tf} = \frac{n_i}{\sum_k n_k} \qquad \mathrm{tfidf} = \mathrm{tf} \cdot \log\left(\frac{|D|}{|(d_j \supset t_i)|}\right)$$

Classes Relevant to the Drama Genre



LOUIE

Menu

GENRE: drama | IMDB

Include fields:

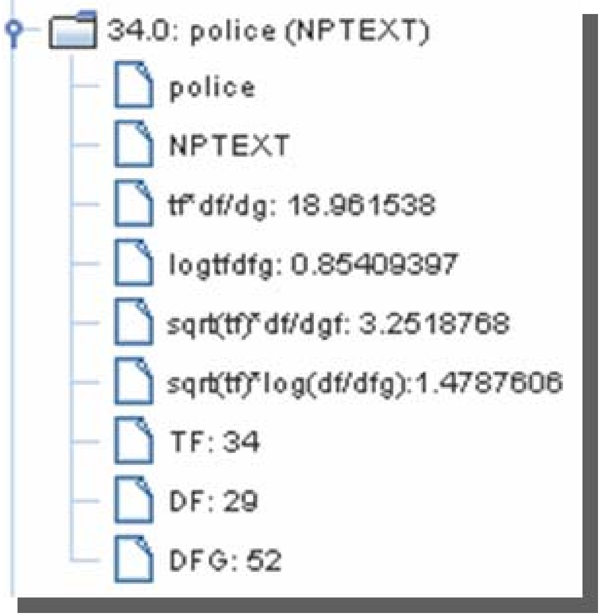
□ Genre □ Keyword □ Director □ Actors □ Title □ Tokens ☑ NP □ Fulltext

List prominent terms | ecexute

```
39.0: murder (NPTEXT)
37.0: night (NPTEXT)
37.0: people (NPTEXT)
36.0: sister (NPTEXT)
34.0: killer (NPTEXT)
34.0: police (NPTEXT)
    police
    NPTEXT
    tf*df/dg: 18.961538
    logtfdfg: 0.85409397
    sqrt(tf)*df/dgf: 3.2518768
    sqrt(tf)*log(df/dfg):1.4787606
    TF: 34
    DF: 29
    DFG: 52
33.0: job (NPTEXT)
33.0: money (NPTEXT)
32.0: group (NPTEXT)
32.0: order (NPTEXT)
32.0: york (NPTEXT)
31.0: ray (NPTEXT)
30.0: agent (NPTEXT)
30.0: dream (NPTEXT)
30.0: jack (NPTEXT)
```

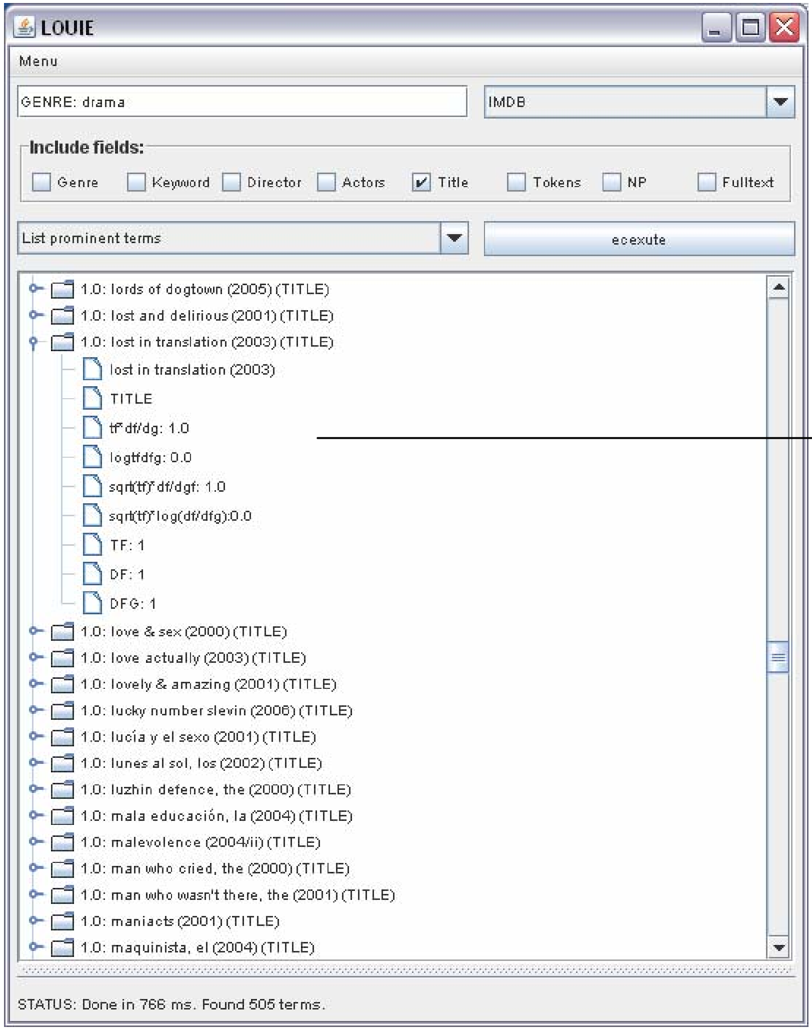
STATUS: Done in 3797 ms. Found 11234 terms.

- Keyphrase extraction technique
- Noun phrases ranked according to various statistical measures



```
34.0: police (NPTEXT)
    police
    NPTEXT
    tf*df/dg: 18.961538
    logtfdfg: 0.85409397
    sqrt(tf)*df/dgf: 3.2518768
    sqrt(tf)*log(df/dfg):1.4787606
    TF: 34
    DF: 29
    DFG: 52
```

Pattern Matching for Learning Individuals



LOUIE

Menu

GENRE: drama

IMDB

Include fields:

☐ Genre ☐ Keyword ☐ Director ☐ Actors ☑ Title ☐ Tokens ☐ NP ☐ Fulltext

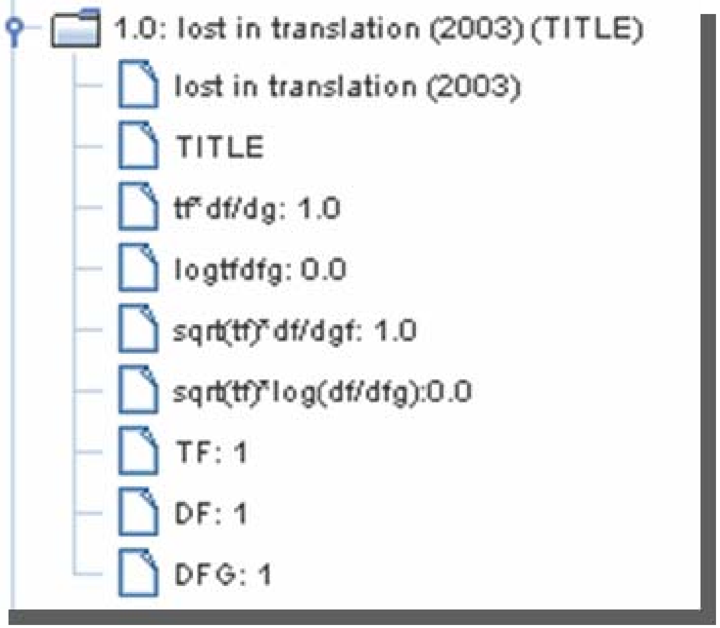
List prominent terms

ecexute

```
1.0: lords of dogtown (2005) (TITLE)
1.0: lost and delirious (2001) (TITLE)
1.0: lost in translation (2003) (TITLE)
    lost in translation (2003)
    TITLE
    tf*df/dg: 1.0
    logtfdfg: 0.0
    sqrt(tf)*df/dgf: 1.0
    sqrt(tf)*log(df/dfg):0.0
    TF: 1
    DF: 1
    DFG: 1
1.0: love & sex (2000) (TITLE)
1.0: love actually (2003) (TITLE)
1.0: lovely & amazing (2001) (TITLE)
1.0: lucky number slevin (2006) (TITLE)
1.0: lucía y el sexo (2001) (TITLE)
1.0: lunes al sol, los (2002) (TITLE)
1.0: luzhin defence, the (2000) (TITLE)
1.0: mala educación, la (2004) (TITLE)
1.0: malevolence (2004/ii) (TITLE)
1.0: man who cried, the (2000) (TITLE)
1.0: man who wasn't there, the (2001) (TITLE)
1.0: maniacts (2001) (TITLE)
1.0: maquinista, el (2004) (TITLE)
```

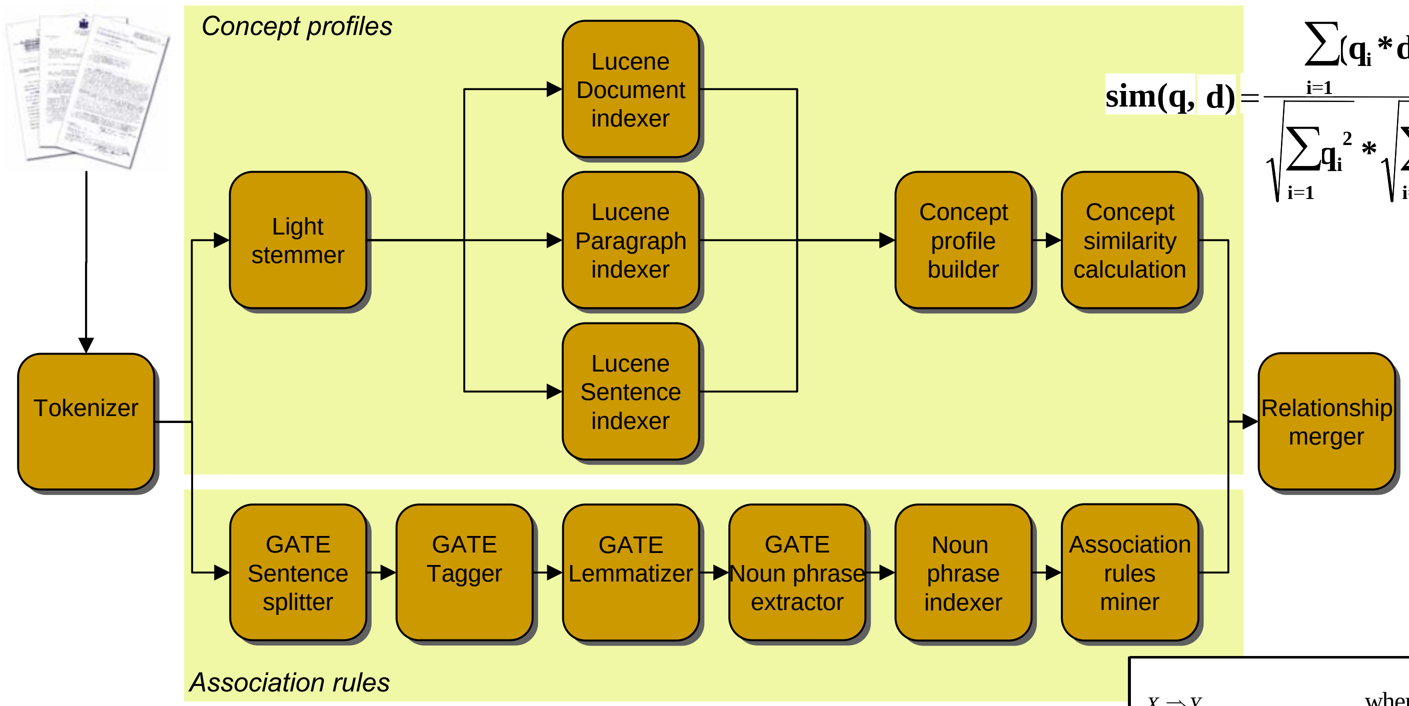
STATUS: Done in 766 ms. Found 505 terms.

- Using structural information (headings, keywords, etc.) to recognize movie instances
- Instances ranked according to various statistical measures



```
1.0: lost in translation (2003) (TITLE)
    lost in translation (2003)
    TITLE
    tf*df/dg: 1.0
    logtfdfg: 0.0
    sqrt(tf)*df/dgf: 1.0
    sqrt(tf)*log(df/dfg):0.0
    TF: 1
    DF: 1
    DFG: 1
```

Learning Relationships (Properties)

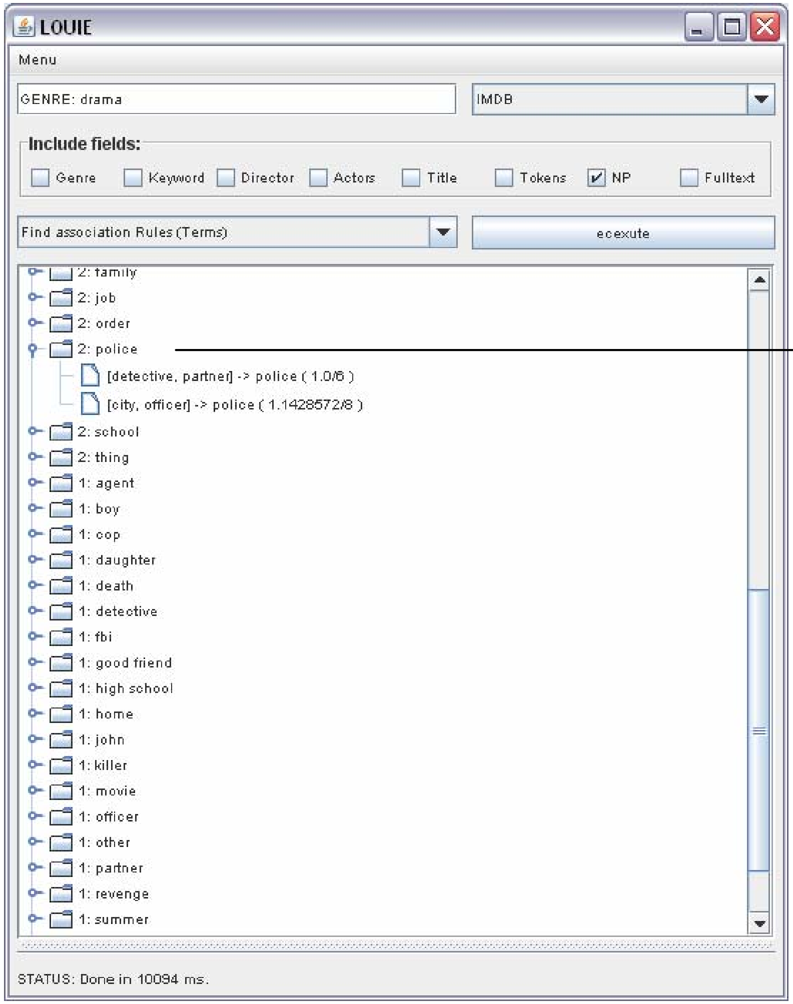


$$\mathbf{sim(q,\ d)} = \frac{\sum_{i=1} (q_i * d_i)}{\sqrt{\sum_{i=1} q_i^{2}} * \sqrt{\sum_{i=1} d_i^{2}}} = \left(\frac{\mathbf{q}}{\|\mathbf{q}\|}\right) \bullet \left(\frac{\mathbf{d}}{\|\mathbf{d}\|}\right)$$

$X \Rightarrow Y$, where $X \subset I, Y \subset I, X \cap Y = Ø$

A rule $X \Rightarrow Y$ holds in the transaction set D with *confidence c* if c% of the transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has *support s* in the transaction set D if s% of the transactions in D contains $X \cup Y$.

Learning Class Relationships



GENRE: drama

IMDB

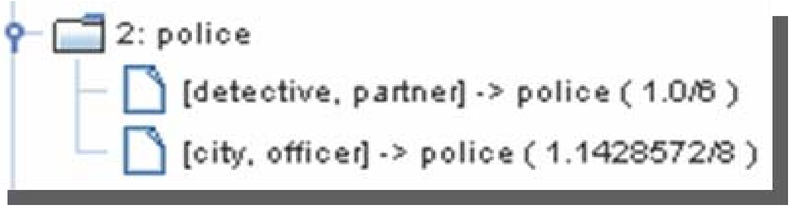
□ Genre □ Keyword □ Director □ Actors □ Title □ Tokens ☑ NP □ Fulltext

Find association Rules (Terms)

```
2: family
2: job
2: order
2: police
    [detective, partner] -> police ( 1.0/6 )
    [city, officer] -> police ( 1.1428572/8 )
2: school
2: thing
1: agent
1: boy
1: cop
1: daughter
1: death
1: detective
1: fbi
1: good friend
1: high school
1: home
1: john
1: killer
1: movie
1: officer
1: other
1: partner
1: revenge
1: summer
```

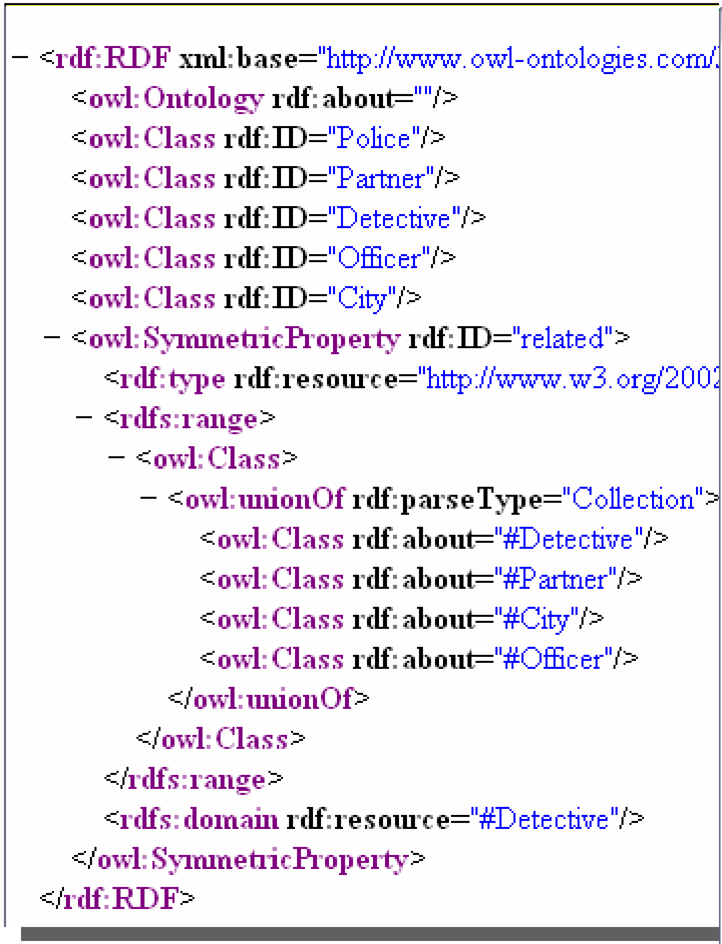
STATUS: Done in 10094 ms.

- Association rules on extracted concepts

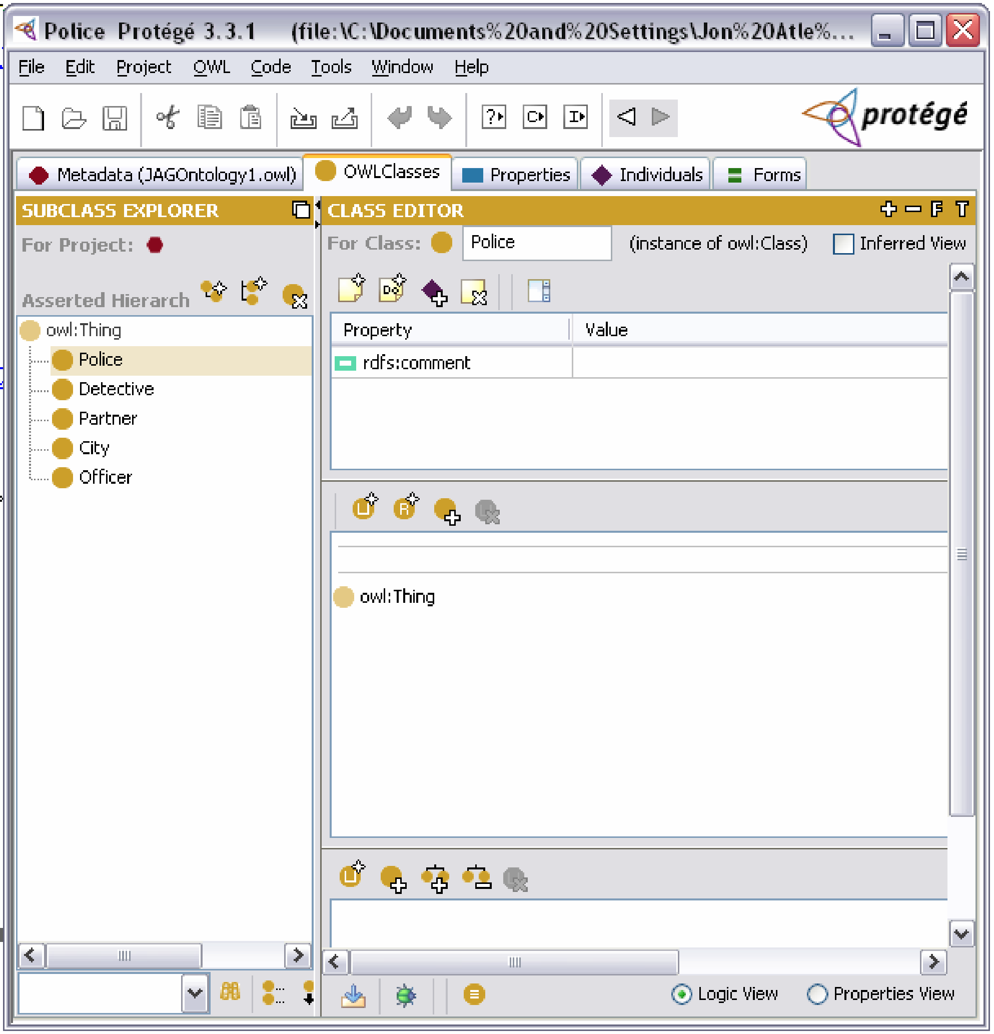


```
2: police
    [detective, partner] -> police ( 1.0/6 )
    [city, officer] -> police ( 1.1428572/8 )
```

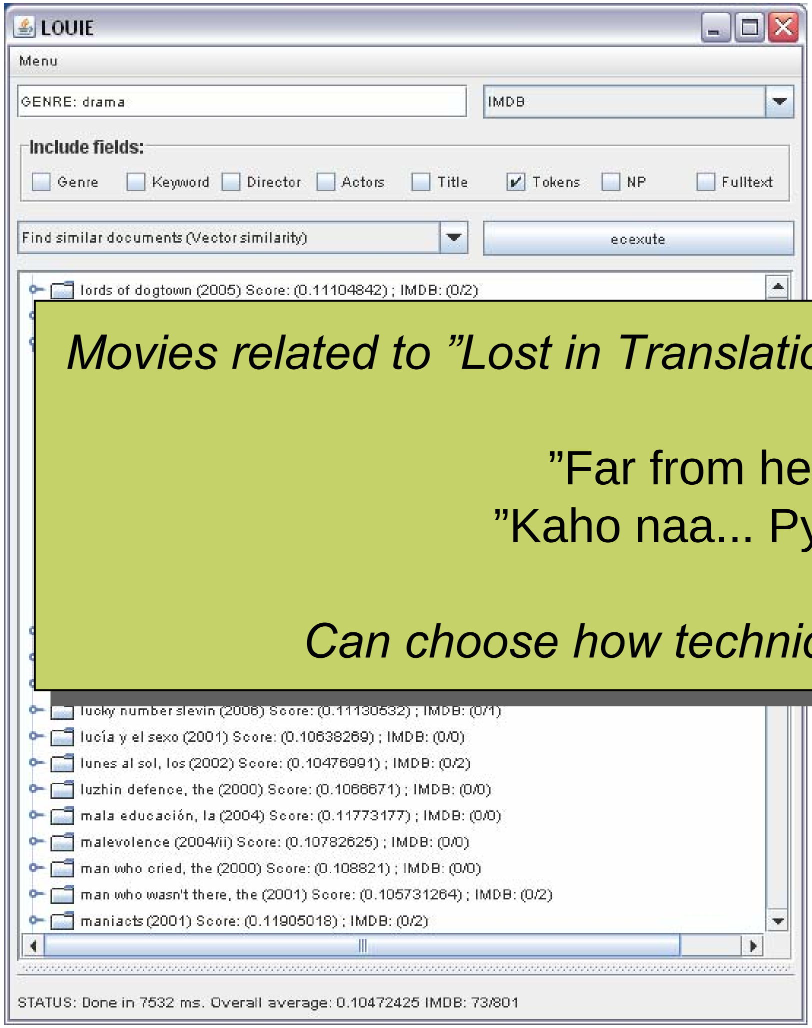
Extract from Police OWL Declaration



```
- <rdf:RDF xml:base="http://www.owl-ontologies.com/
    <owl:Ontology rdf:about=""/>
    <owl:Class rdf:ID="Police"/>
    <owl:Class rdf:ID="Partner"/>
    <owl:Class rdf:ID="Detective"/>
    <owl:Class rdf:ID="Officer"/>
    <owl:Class rdf:ID="City"/>
  - <owl:SymmetricProperty rdf:ID="related">
      <rdf:type rdf:resource="http://www.w3.org/2002
    - <rdfs:range>
      - <owl:Class>
        - <owl:unionOf rdf:parseType="Collection">
            <owl:Class rdf:about="#Detective"/>
            <owl:Class rdf:about="#Partner"/>
            <owl:Class rdf:about="#City"/>
            <owl:Class rdf:about="#Officer"/>
          </owl:unionOf>
        </owl:Class>
      </rdfs:range>
      <rdfs:domain rdf:resource="#Detective"/>
    </owl:SymmetricProperty>
  </rdf:RDF>
```



Learning Relationships between Movies



GENRE: drama

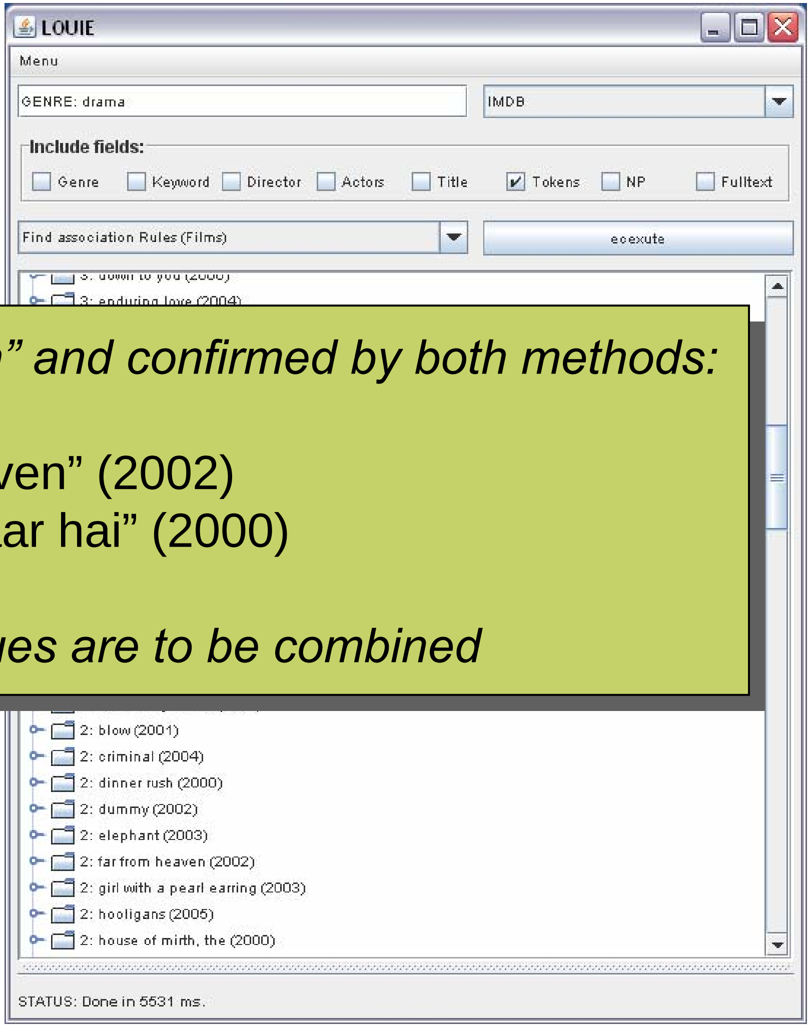
□ Genre □ Keyword □ Director □ Actors □ Title ☑ Tokens □ NP □ Fulltext

Find similar documents (Vector similarity)

```
lords of dogtown (2005) Score: (0.11104842) ; IMDB: (0/2)
lucky number slevin (2006) Score: (0.11130632) ; IMDB: (0/1)
lucía y el sexo (2001) Score: (0.10638269) ; IMDB: (0/0)
lunes al sol, los (2002) Score: (0.10476991) ; IMDB: (0/2)
luzhin defence, the (2000) Score: (0.1066671) ; IMDB: (0/0)
mala educación, la (2004) Score: (0.11773177) ; IMDB: (0/0)
malevolence (2004/ii) Score: (0.10782625) ; IMDB: (0/0)
man who cried, the (2000) Score: (0.108821) ; IMDB: (0/0)
man who wasn't there, the (2001) Score: (0.105731264) ; IMDB: (0/2)
maniacts (2001) Score: (0.11905018) ; IMDB: (0/2)
```

STATUS: Done in 7532 ms. Overall average: 0.10472425 IMDB: 73/801

Concept vector similarities



GENRE: drama

□ Genre □ Keyword □ Director □ Actors □ Title ☑ Tokens □ NP □ Fulltext

Find association Rules (Films)

```
3: down to you (2000)
3: enduring love (2004)
2: blow (2001)
2: criminal (2004)
2: dinner rush (2000)
2: dummy (2002)
2: elephant (2003)
2: far from heaven (2002)
2: girl with a pearl earring (2003)
2: hooligans (2005)
2: house of mirth, the (2000)
```

STATUS: Done in 5531 ms.

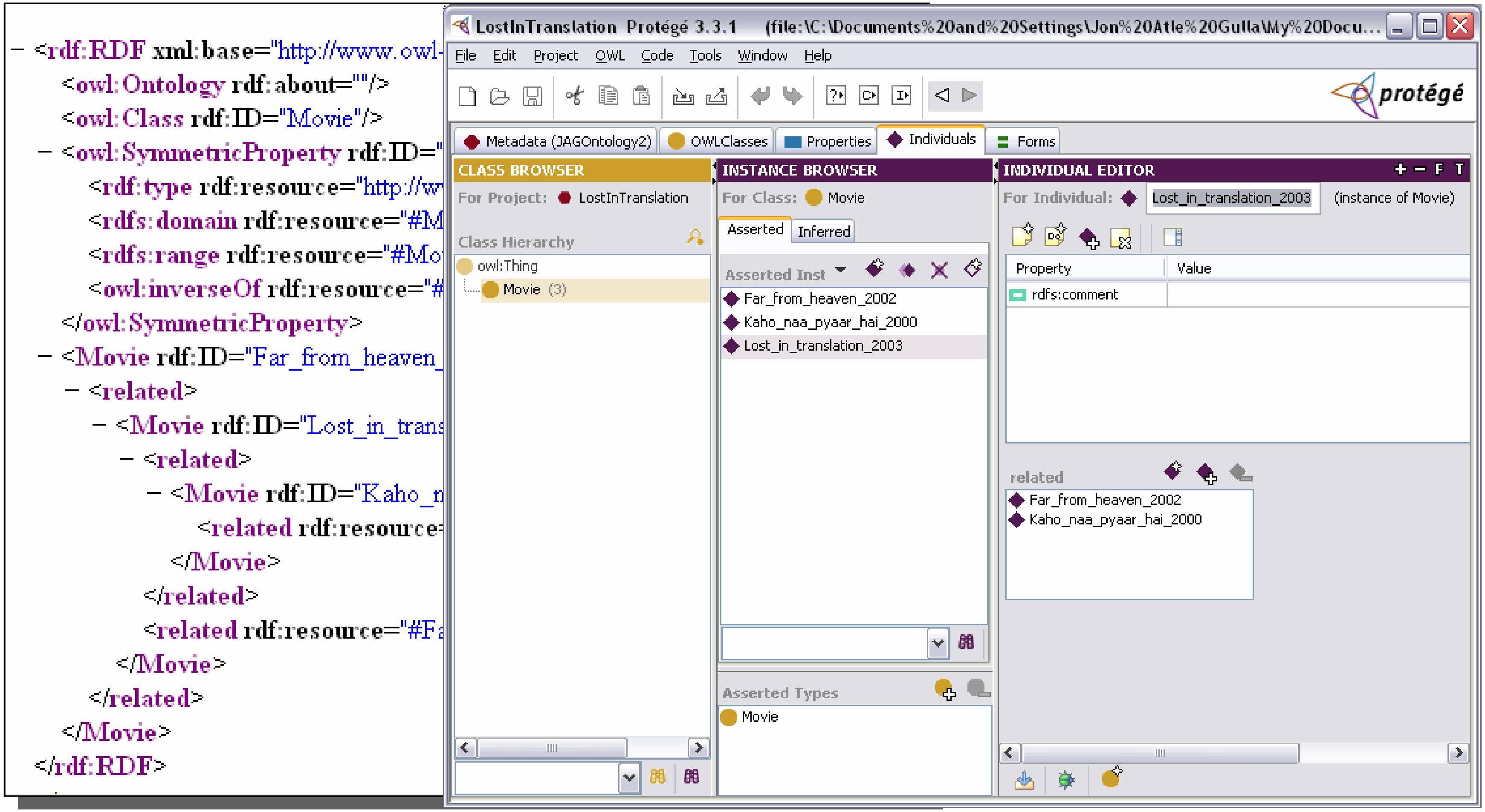
Association rules

Movies related to "Lost in Translation" and confirmed by both methods:

"Far from heaven" (2002)
"Kaho naa... Pyaar hai" (2000)

Can choose how techniques are to be combined

Extract from OWL Generation



```
– <rdf:RDF xml:base="http://www.owl-
    <owl:Ontology rdf:about=""/>
    <owl:Class rdf:ID="Movie"/>
  – <owl:SymmetricProperty rdf:ID="
      <rdf:type rdf:resource="http://w
      <rdfs:domain rdf:resource="#M
      <rdfs:range rdf:resource="#Mo
      <owl:inverseOf rdf:resource="#
    </owl:SymmetricProperty>
  – <Movie rdf:ID="Far_from_heaven_
    – <related>
      – <Movie rdf:ID="Lost_in_trans
        – <related>
          – <Movie rdf:ID="Kaho_n
              <related rdf:resource=
            </Movie>
          </related>
          <related rdf:resource="#F
        </Movie>
      </related>
    </Movie>
  </rdf:RDF>
```

For Project: LostInTranslation

owl:Thing
Movie (3)

For Class: Movie

Far_from_heaven_2002
Kaho_naa_pyaar_hai_2000
Lost_in_translation_2003

Asserted Types: Movie

For Individual: Lost_in_translation_2003 (instance of Movie)

Property	Value
rdfs:comment	

related
Far_from_heaven_2002
Kaho_naa_pyaar_hai_2000

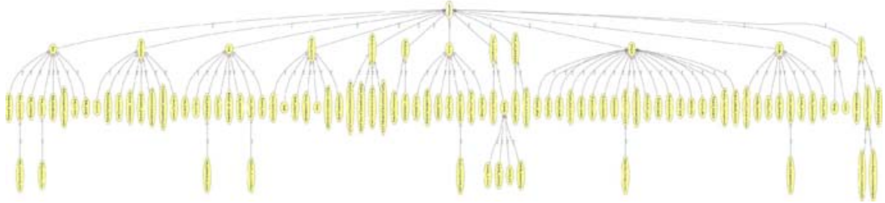
Quality of Class Learning

Evaluation Procedure

- Extracted candidates from project management domain (PMBOK):
  - 50,600 tokens (ca. 130 pages)
  - Generated candidates for each area (chapter)
- Constructed ontology from candidates (with help from STATOIL employee)
- Built an alternative ontology manually (with help from another STATOIL employee)
- Compared quality of two ontologies for domain representation
- (Compared quality of two ontologies in ontology-driven (semantic) search)

Results for Class Learning Evaluation

- Domain representation:



Semi-automatically constructed ontology for project management

	Classes	Hierarchical levels	Very good classes	Acceptable classes
Semi-automatic	106	3	73 (79%)	33 (21%)
Manual	142	5	122 (86%)	20 (14%)

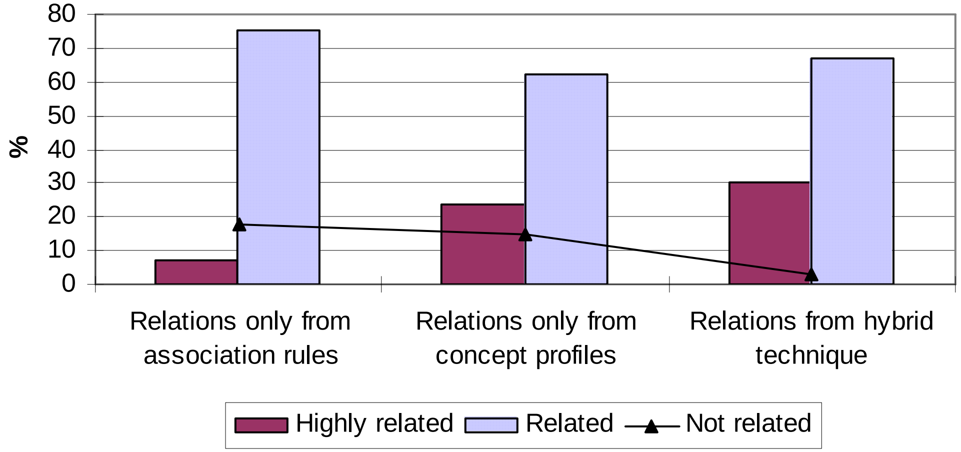
- 62 (58% of semi-automatic ontology) classes identical
- Tool-generated ontology:
  - Slightly smaller, with less abstraction levels
  - Almost as good as manually built ontology
  - Substantially faster to build
  - Easy to improve further

Semi-automatic ontology construction very promising!

Quality of Relationship Learning

- Experiment with Statoil’s project management standard (PMI)
  - Generated class relationships based on PMBOK
  - Quality of relationships verified by project management experts
  - Comparison between association rules and concept vector similarity

- Result of evaluation



Conclusions

- Ontology Learning is the discipline of automatically or semi-automatically constructing ontologies
- Challenge to construct and maintain search ontologies
- Numerous learning strategies
  - Classes
  - Individuals
  - Relationships (properties)
- Ontology learning produces an intial fragmentary OWL model
  - Manual verification and correction
  - Manual completion of missing parts
  - But: Quality of techniques improving
- *Ontology learning a complement to traditional ontology engineering methodologies*