Towards Retrieving Scholarly Literature via Ontological Relationships

Vojtěch Svátek, Ondřej Šváb

Katedra informačního a znalostního inženýrství
VŠE Praha
http://keg.vse.cz

Agenda

- Searching scholarly literature: problem formulation
- Inadequacy of existing approaches
- Generic proposal for a new (combined) approach
- Very small initial experiment
- Future plans and discussion

Searching scholarly literature…

- Frequent use, e.g.
  - PhD student who wants to compare his/her research idea with the state-of-the-art
  - Researcher needing material for the ‘related work’ section of a paper
  - Business intelligence worker wanting to carry out ‘technological watch’ wrt. potential innovations in the field of interest for the company

… as accessing the space of ideas

- Historically, the science inevitably evolved around individual ‘schools of thought’
  - The contacts among different ‘schools of thought’ as well as between academia and industries were only scarce (paper documents, intermittent oral communication…)
- Recently, the availability of documents on the WWW (and web-accessible digital libraries) made the old barriers disappear in many domains
  - A researcher or practitioner can in principle instantly retrieve publications by various ‘schools of thought’, even beyond a single problem domain
- But does the existing search technology promise relevant results?

Does web/DL search fit?

- Current open web search tools index even (at least abstracts of) scholarly papers in digital libraries (DL)
- However, their retrieval bias is unsuitable for this finding specialised literature
  - PageRank strongly prefers often-cited documents, which usually deal with generic topics
  - For obtaining relevant results on specific topics, one mostly needs to introduce specific terms
  - Unfortunately, specific terms are not only specific for a problem at question but also for a certain ‘school of thought’, i.e. relevant publications by another ‘school’ are cut off
  - Moreover, sometimes, rather than the terms themselves, it is the way methods, tools, resources etc. are related to each other what matters!

Simple example

- Information request:
  “Has anyone used a statistical information extraction tool in order to acquire background knowledge from Wikipedia, which will in turn be used within an adaptive e-learning system? Or something similar?”
- Querying Google (or another engine) just using terms like *“statistical”*, *“information extraction”*, *“background knowledge”*, *“Wikipedia”*, *“e-learning”* will invariably lead to ambiguous results

Common approaches to solving this problem (1)

- 'Heavy-weighted' semantic web: annotating documents according to concepts from carefully-designed domain ontologies
  - potentially captures complex semantics of the content and thus allows for very precise querying
- However
  - manual annotation does not scale due to the high cost of explaining the ontology to the annotators (who actually have to be domain experts)
  - NLP-based annotation quite erroneous due to 'semantic gap' between extractable concepts and high-level ontological representation
  - dependence on a pre-defined domain-specific ontology challenges the possibility of cross-domain search

Common approaches to solving this problem (2)

- Folksonomy-based approaches (e.g. BibSonomy): annotating documents with arbitrary tags
  - 'democratises' the annotation task wrt. crowds of volunteers, thus significantly increasing the annotation base
  - there are no true cross-domain boundaries
- However
  - disambiguation of isolated ad hoc tags is hard, as there is no other clue than statistical co-occurrence
  - even with correct disambiguation, there is no account for relationships among concepts in the context of a given publication

Proposed approach

- Lightweighted relational representation
  - typed entities and n-ary relationships with roles
  - conversion to/from RDF, Topic Maps and possibly other formalisms
- Bottom-up construction of conceptual structures in this representation
  - allowing for annotation by ad hoc relational tags by volunteers (a la folksonomy)
  - but support by NLP-based content analysis, similarity-based recommendation and (where available) simplified domain ontologies
- Conceptual structure merging
  - across different annotators
  - merging patterns are empirically discovered for future use
- Conceptual structure semantic interpretation
  - especially via alignment of the merged structures with existing ontology content design patterns

Progress of the research

- 99,5%: future work
- 0,5%: will be presented now

Concept of experiment

- Independent creation of relational annotations by different annotators for the same publications
- Comparison of the annotations, plus computation of simple statistics
- Formulation of sample merging patterns

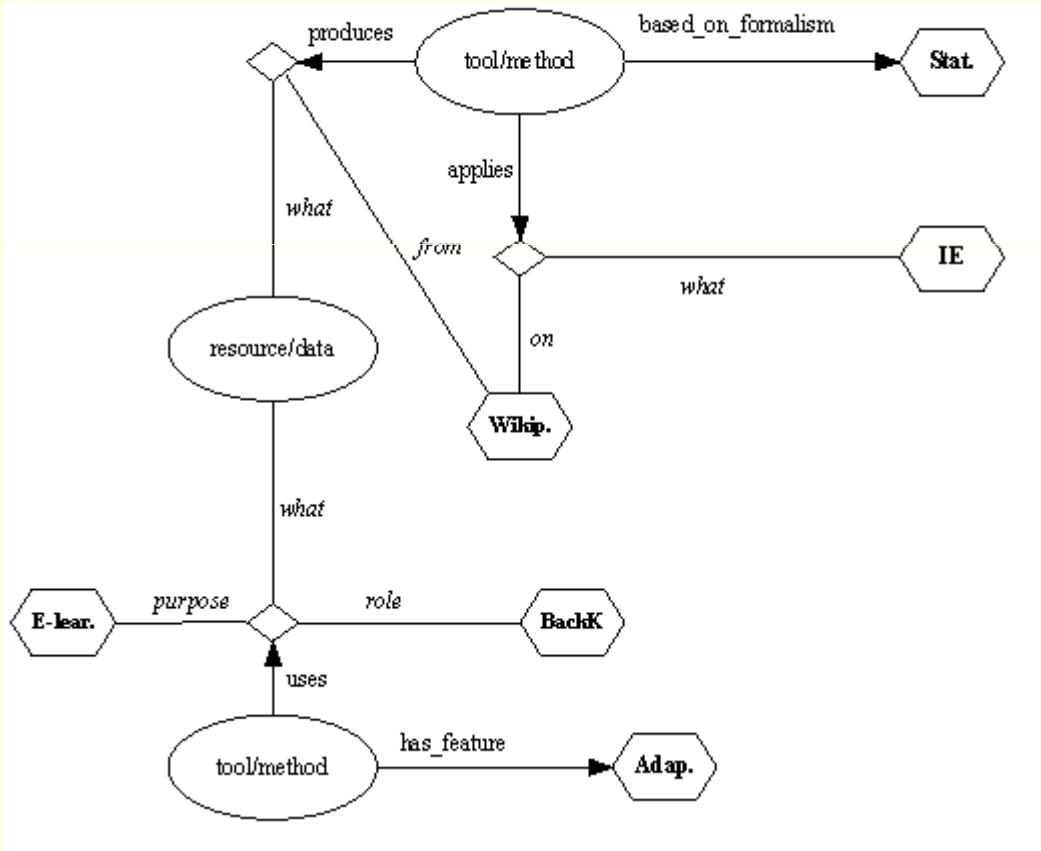
Specific settings

- Two annotators, five publications to be annotated
  - annotators = authors of this paper
- Extremely small seed vocabulary
  - just what was contained in the single, previously mentioned illustrative query!
- Non-binding verbal guidelines for annotation
  - recommendation: 10-20 relations per publication
  - reuse of seed vocabulary where natural, but introduction of new entities wherever needed
  - bias towards use of relations over concepts (esp. over relation reifications) where possible

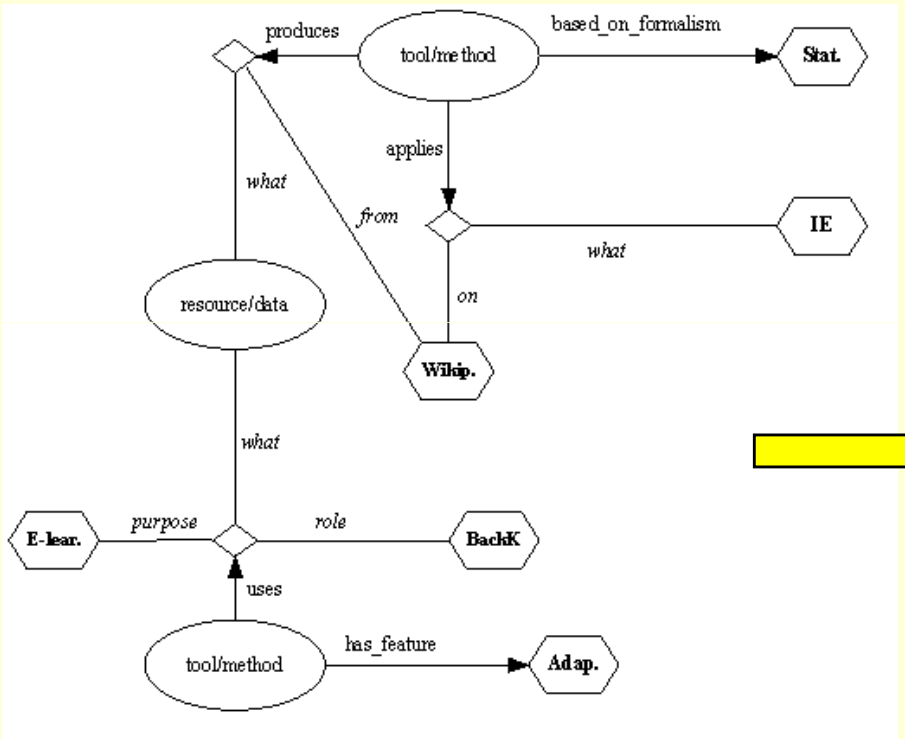
Seed vocabulary query

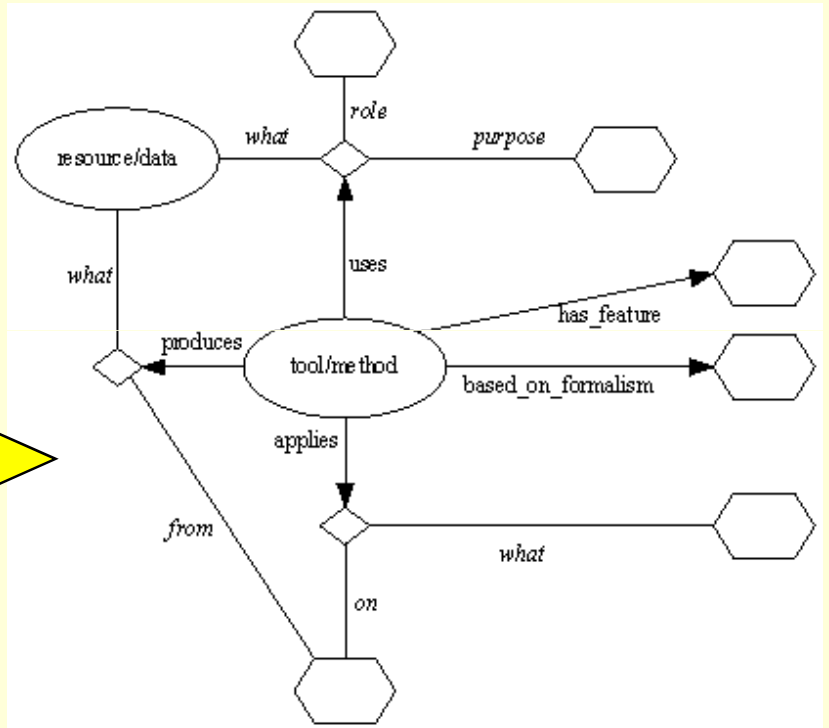
"Has anyone used a statistical information extraction tool in order to acquire background knowledge from Wikipedia, which will in turn be used within an adaptive e-learning system?"

TOOL1 has_type tool/method
TOOL1 based_on_formalism
    Statistics
TOOL1 applies
    what:Information_extraction
    on:Wikipedia
TOOL1 produces
    what:RESOURCE1
    from:Wikipedia
RESOURCE1 has_type
    resource/data
TOOL2 has_type tool/method
TOOL2 has_feature Adaptivity
TOOL2 uses what:RESOURCE1
    purpose:E-learning
    role:Background_knowledge

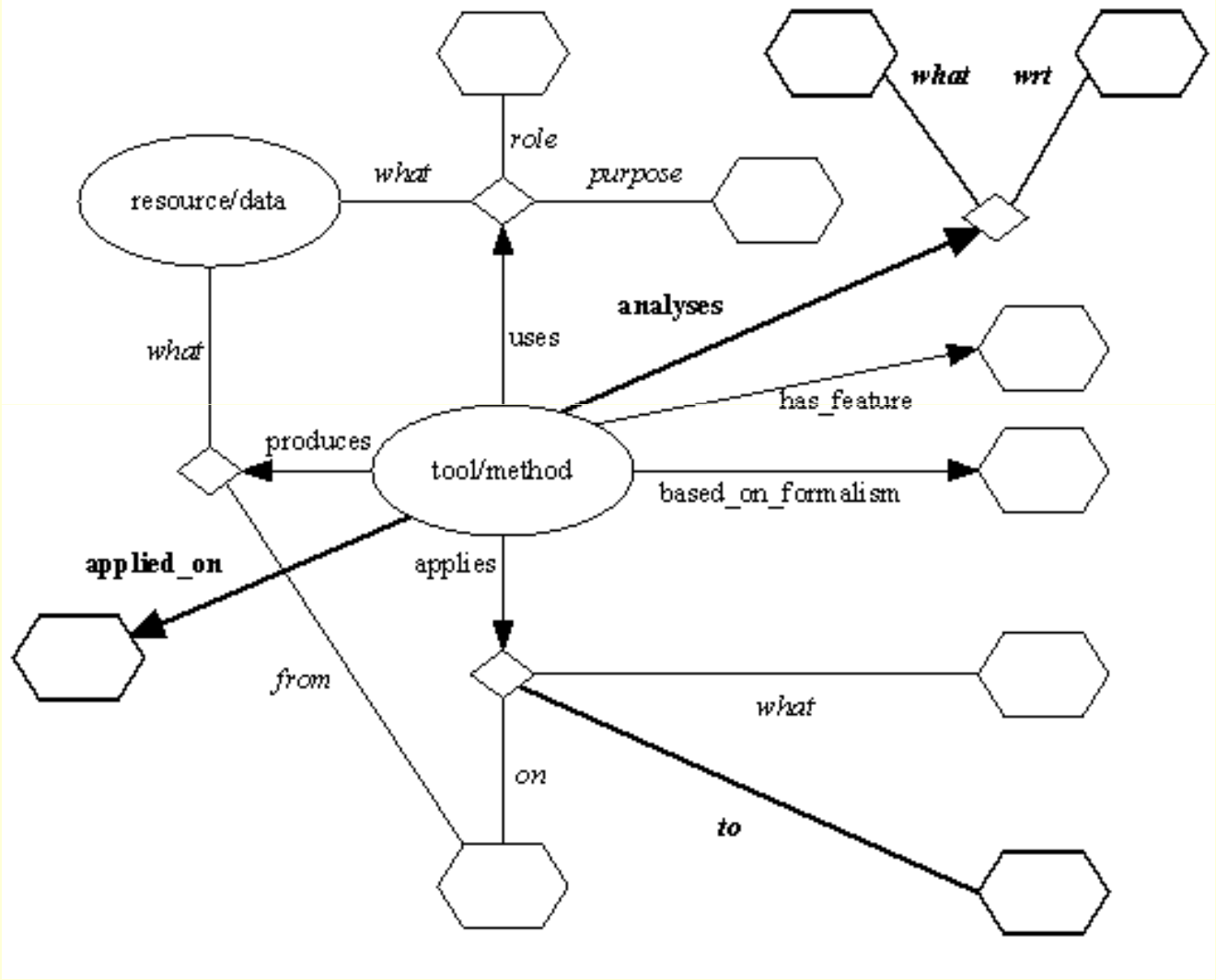


Abstracted seed vocabulary





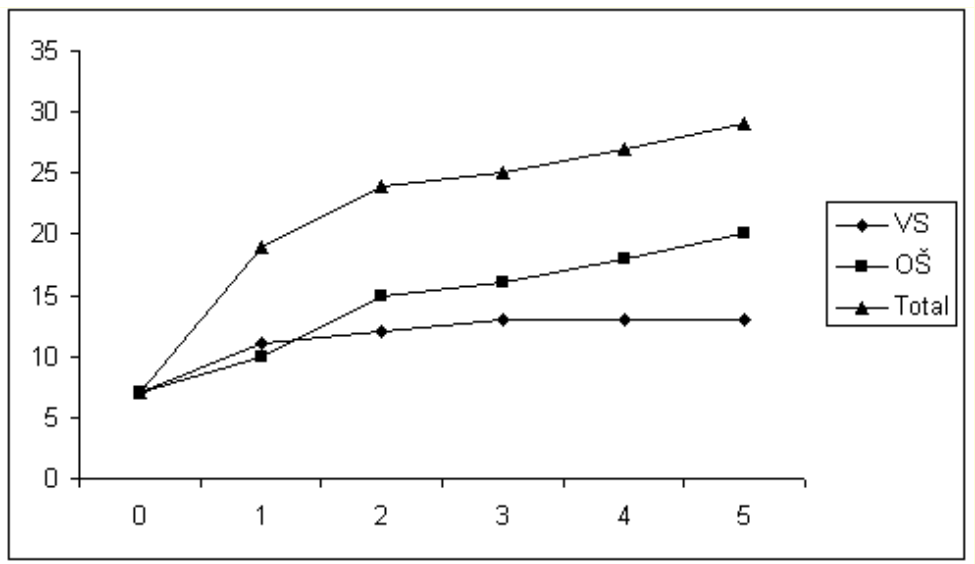
Evolution of vocabulary



Evolution statistics

Annotator	VS	OŠ
Max. no. of relation instances	13	22
Min. no. of relation instances	5	8
Avg. no. of relation instances	7.2	13.0
New relations introduced	6	9
New types introduced	0	4

Number of relations + types



Pattern discovery example

- Annotator 1
  - METHOD1 has_type tool/method
  - METHOD1 produces what:Ontology from:Source_code
- Annotator 2
  - TOOL2 has_type tool
  - TOOL2 applies what:code_analysis on:source_code
- Possible conclusion
  - Assuming X ∈ tool/method often co-occurs with X ∈ tool*:*
    X applies what : Y on : Z
    could often co-occur with
    X produces what : W from : Z

Most imminent future steps

- Put the suggested relational annotation language on a formal basis
  - including transformation to/from RDF and TM
- Investigate instant gratification approaches in volunteer-based annotation
- Investigate the conceptual structure matching and merging pattern discovery methods, plus the possibility of alignment with ontology content patterns

Thank you!
Discussion?!