Informal vs. Formal Semantic Approaches to Document Content

Jon Atle Gulla
Norwegian University of Science and Technology, Trondheim, Norway
Email: jag@idi.ntnu.no

Manual vs. automatic processes?

Authorative vs. social processes?

Content of Unstructured Documents

- Current features:
  - Retrieval
  - Categorization and clustering
  - Summarization
  - Document taxonomies
  - Sentiment
- Research:
  - Question-answering
  - Reasoning about document content
  - Opinions

Unstructured Document Content

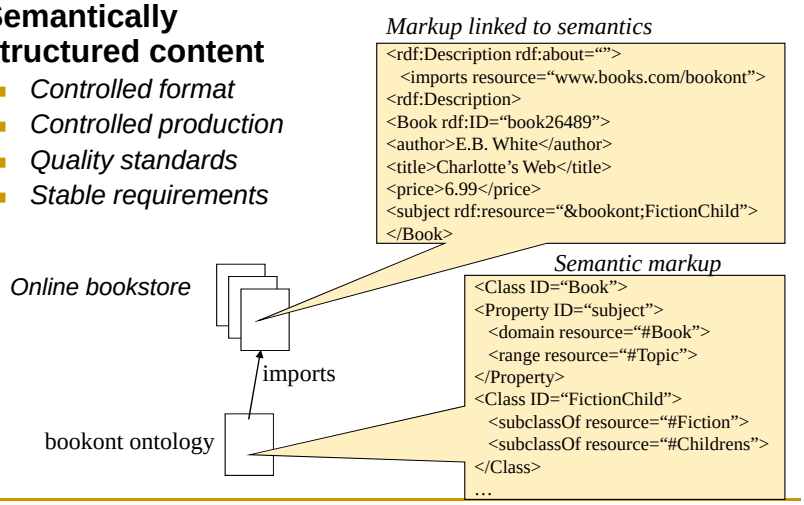
Features of Unstructured Document Repositories

- Continuous evolution of repository
- Uncontrolled production
- Multitude of producers
- Variable quality

- *Semantics useful only if semantic representations aligned with document content*
  - *Concepts*
  - *Relationships*

Not Focus of Talk…

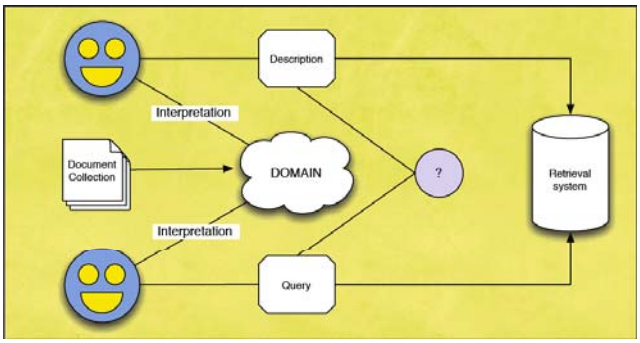
- **Semantically structured content**
  - *Controlled format*
  - *Controlled production*
  - *Quality standards*
  - *Stable requirements*



Why Semantics for Unstructured Document Repositories?

The Language Problem

- People may use different words to refer to the same phenomena



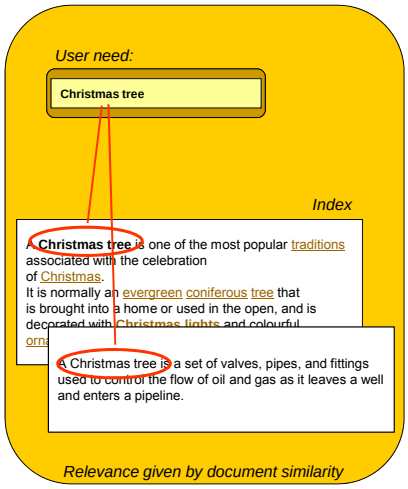
Terminologies are challenging

- Example: Medical diagnosis

Patient communication	Collegial Discussion	Reports & Documentation	Latin
Blindtarmbetennelse (Appendicitis)	Akutt Appendicitt	Akutt Appendicitt	Appendicits Acuta
Halsbetennelse (Tonsilitis)	Tonsilitt	Akutt Tonsilitt	Tonisilitis Acuta
Lårhalsbrudd (Broken hip)	Collumfraktur / Kollifem	Fractura Colli Femoris	Fractura Colli Femoris
Kronisk Bronkitt (Chronic broncitis)	KOLS	Kronisk Obstruktiv Lungesykdom	
Forstoppelse (Congestion)	Obstipasjon	Obstipasjon	Obstipatio
Prostatakreft (Prostate cancer)	Cancer Prostata	Cancer Prostata	
Høysnue (Hay fever)	Allergisk Rhinoconjunctivitt	Allergisk Rhinoconjunctivitt	

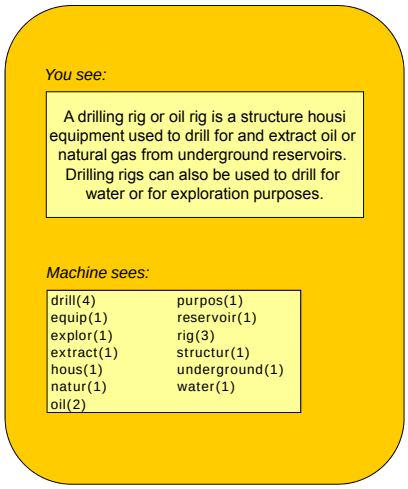
Traditional Search Principles

- Bag-of-words principle
  - Machine understands document as a set of word frequencies
- Word matching principle
  - *Syntactic search:* Relevant documents are documents that contain exactly those words that appear in the query
  - *Morpho-syntactic search:* Relevant documents are documents that contain inflectional variants of exactly those words that appear in the query
- One shot principle
  - Query and result set ignored when new query is posted



Traditional Search Principles

- Bag-of-words principle
  - Machine understands document as a set of word frequencies
- Word matching principle
  - *Syntactic search:* Relevant documents are documents that contain exactly those words that appear in the query
  - *Morpho-syntactic search:* Relevant documents are documents that contain inflectional variants of exactly those words that appear in the query
- One shot principle
  - Query and result set ignored when new query is posted



Traditional Search Principles

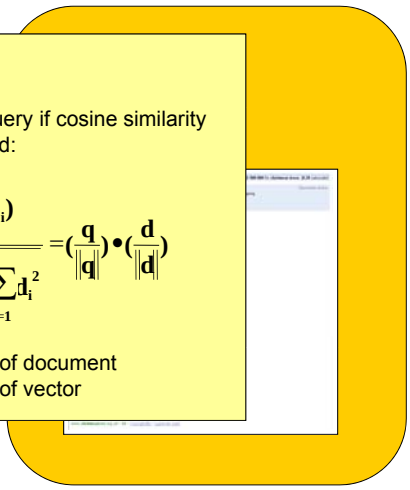
- Bag-of-words
  - Machine un
    as a set of w
- Word matchi
  - *Syntactic se*
    Relevant do
    documents
    those words
    query
  - *Morpho-syn*
    Relevant do
    documents
    inflectional
    those words
    query
- One shot pri
  - Query and result set ignored when new query is posted

Implementation:

Document relevant to query if cosine similarity above a certain threshold:

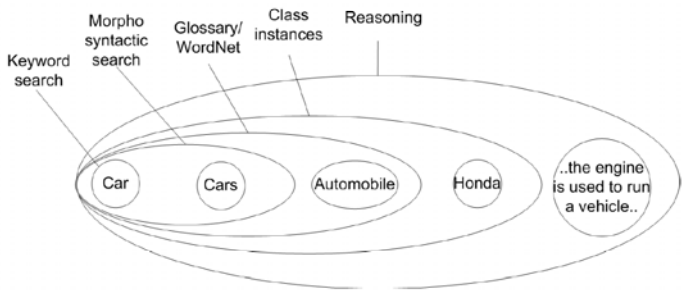
$$sim(q, d) = \frac{\sum_{i=1}(q_i * d_i)}{\sqrt{\sum_{i=1} q_i^2} * \sqrt{\sum_{i=1} d_i^2}} = (\frac{q}{\|q\|}) \bullet (\frac{d}{\|d\|})$$

d: vector representation of document
q: vector representation of vector



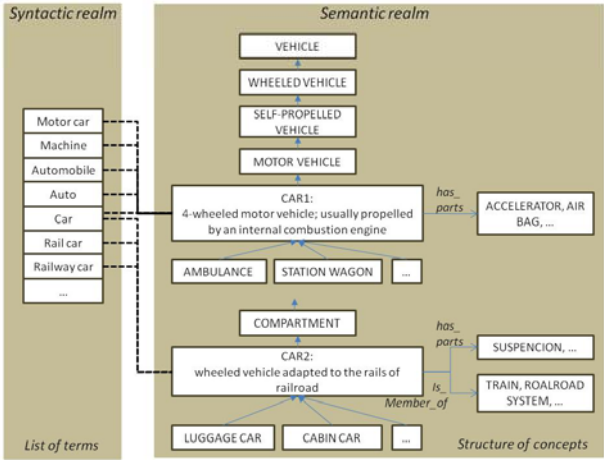
Levels of Intelligent Search

- *Which documents should be returned for the query 'car'*



Ontologies should help us see the semantic relationships between words

Terms are not Concepts



WordNet

From Syntactic to Semantic Search

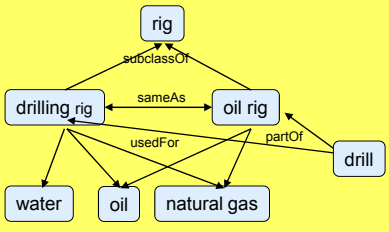
"A drilling rig or oil rig is a structure housing equipment used to drill for and extract oil or natural gas from underground reservoirs. Drilling rigs can also be used to drill for water or for exploration purposes." (Ref: Wikipedia)

Traditional Search Principle:

drill(4)	purpos(1)
equip(1)	reservoir(1)
explor(1)	rig(3)
extract(1)	structur(1)
hous(1)	underground(1)
natur(1)	water(1)
oil(2)	

Document text is just a set of strings

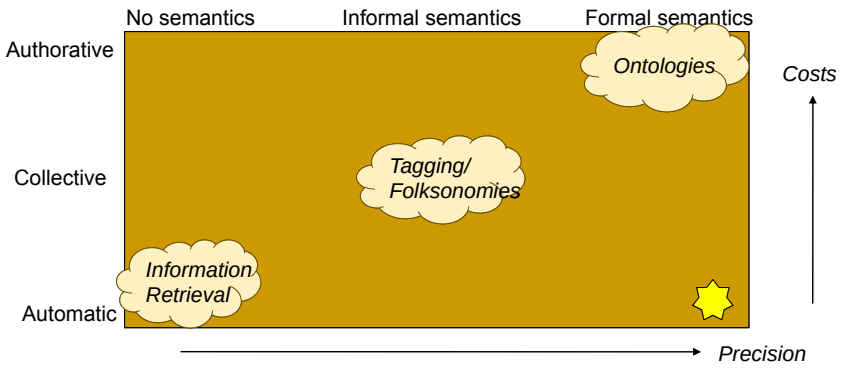
Semantic Search Principle:



Use semantics to represent domain vocabulary, documents' content and/or user's information needs

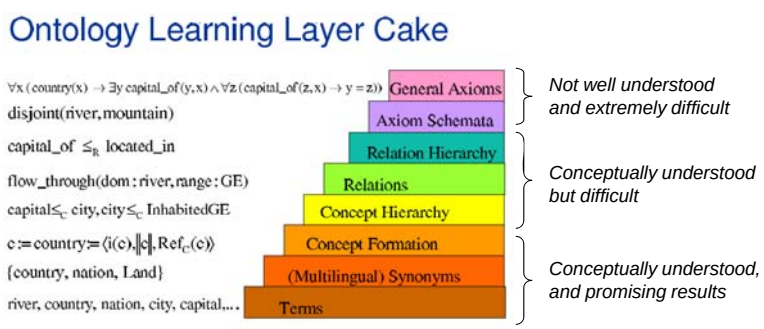
Levels of Semantics

Semantics in Content Management



Ontology Rich in Structure

- … but expensive to build and maintain

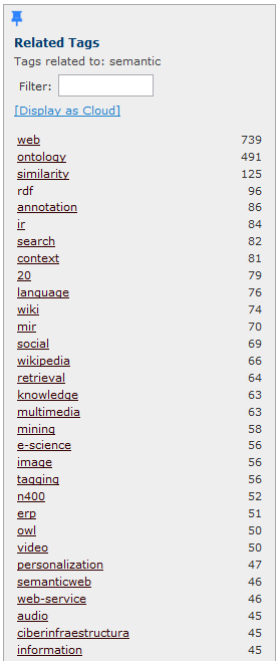


Folksonomies Cheap and Dynamic

- … but lack proper structures/ hierarchies

www.citeulike.org

Site for managing, discovering and sharing schoolarly references



What Statistics and Linguistics can do to Ontologies

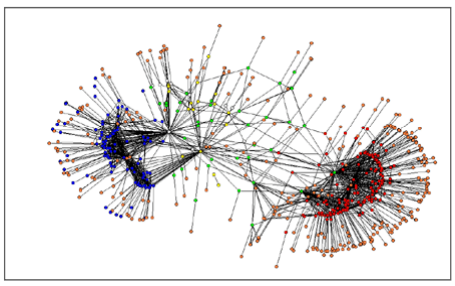
- Continuous evolution of repository
- Uncontrolled production
- Multitude of producers
- Variable quality

- *Semantics useful only if semantic representations aligned with document content*
  - *Concepts*
  - *Relationships*

Ontology Learning

- A collection of techniques for proposing candidate classes, individuals, properties, etc. in ontologies
- Statistical and linguistic aproaches
- Supporting ontology engineers, not replacing them

- *Variable quality, but speeds up modeling process*



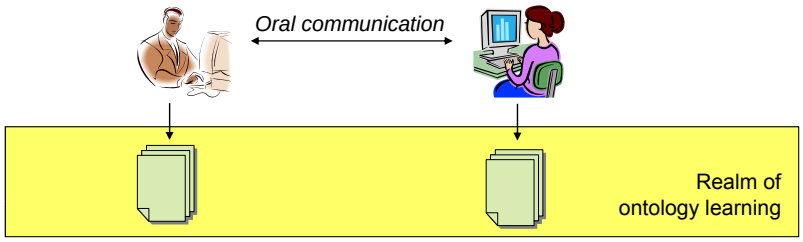
Ontology Modeling vs. Learning

- Traditional ontology engineering approach
  - *Project:* Form team of ontology and domain experts
  - *Ontology & domain experts:* Model domain with concepts and relationships
  - *Domain experts:* Test ontology against domain knowledge
  - *Ontology experts*: Verify internal quality and application quality
- Expensive and time-consuming approach

- Ontology learning approach:
  - *Domain experts:* Find representative domain text
  - *Tool:* Extract candidate concepts and relationships automatically from domain texts
  - *Ontology & domain experts:* Select candidates and relationships and complete model
- Can also be used to verify domain quality of existing ontology
- Very cost-effective approach

Ontology Learning Assumptions

- People communicate using domain-specific concepts
- People document using domain-specific concepts
- Ontology learning makes use of written documentation rather than human involvement

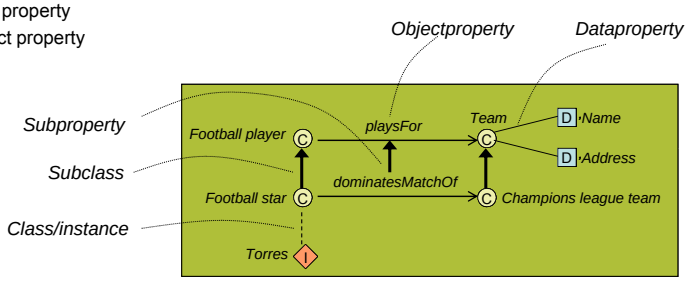


- Requirements:
  - Documents representative for domain terminology
  - Documents cover all the terminology
  - Well-defined and consistent use of terminology in domain

Example: Ontology Relationships

Types of relationships in OWL DL

- Taxonomic relationships
  - Class/individual
  - Subclass
  - Subproperty
- Non-taxomomic relationships
  - Data property
  - Object property



Association Rules

- Identification of term relationships on the basis of co-occurrence data
- Formal definition:

D: Document collection
T. Set of documents in collection
I: Set of terms

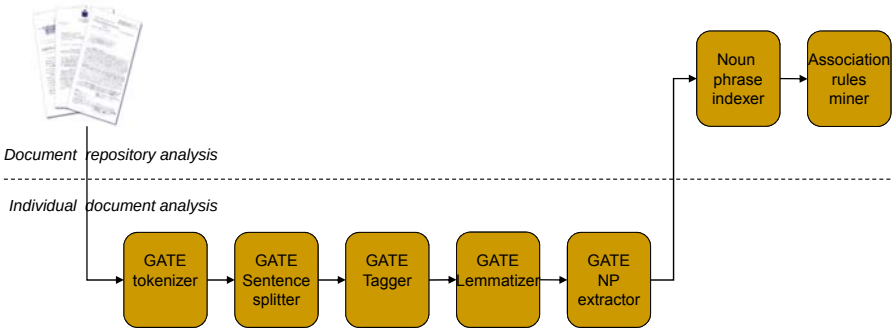
Association rule: $X \Rightarrow Y$, where $X \subset I, Y \subset I, X \cap Y = \varnothing$

Confidence c: c% of documents T that contain X also contain Y
Support s: s% of documents in collection contain $X \cup Y$

- Example:
  {James Bond, young woman} $\Rightarrow$ seduction

Learning Process

- Using GATE for the linguistic pre-processing stage
- Using Lucene for NP indexing
- Associaton rule miner developed in Java



Alternative Approach

Concept Similarity Calculations

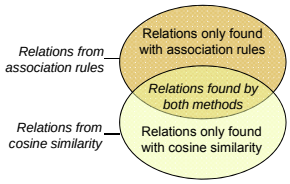
- Compute term vector for each concept in ontology
  - Words included in the same documents included in vector
  - Weights reflect
    - Frequency of word
    - Proximity to concept references in the documents
- Calculate cosine similarity between concepts

$$cos(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Assume relationship if similarity above predefined threshold

Statoil Case

- Domain:
  - PMI methodology used by STATOIL and other large companies
  - PMBOK: 50.600 tokens – 12 chapters
- Extracted relatinships for both methods
  - Association rules
  - Cosine similarity method
- External domain experts validated each relationship
  - Highly related
  - Somewhat related
  - Not related
- Comparative evaluation
  - Three sets of results



Categories for comparative evaluation

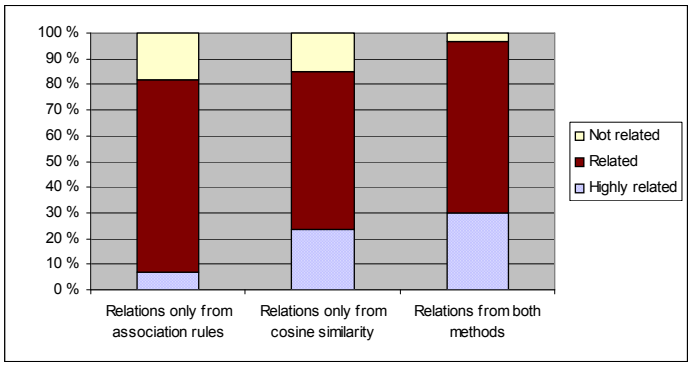
Extracted Relationships

Terms related to the concept Cost

Related concepts only from assocation rules		Related concepts only from cosine similarity		Related concepts from both methods	
project management team	R	cost management	HR	activity	R
management team	R	cost baseline	HR	assumption	NR
organization	R	actual cost	HR	control	R
product	HR	schedule	R	cost estimate	HR
information	R	project schedule	R	performance	R
tool	R	earn value	R	process	R
project team	R	staff	R	project	HR
application area	R	project staff	R	project management	R
risk analysis	R	milestone	NR	project objective	R
result	R	plan value	R	project plan	R
risk	R	stakeholder	HR	quality	R
resource	R	project deliverable	R	scope	R
consequence	R	ev	NR	scope statement	R
estimate	R	earn value management	R		
phase	NR	management	R		
probability	R	scope definition	NR		
action	R	scope management	R		
analysis	R	customer	R		
seller HR	HR	sponsor	R		
		project management IS	R		
		constraint	R		
		project manager	R		
		project plan development	R		
		procurement management	NR		
		project plan execution	NR		
		quality management	R		
		work breakdown structure	R		

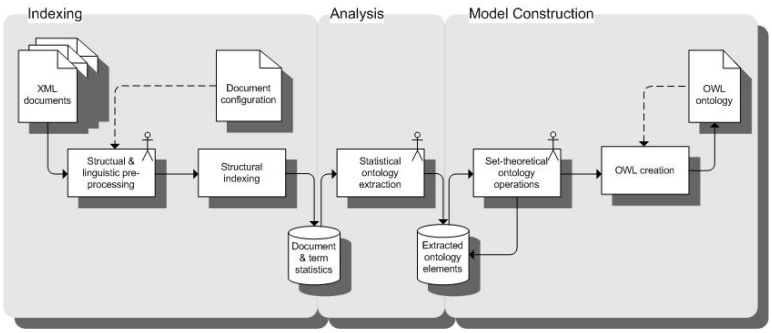
Evaluation

- Cosine similarity method extracted more specialized relationships than association rules
- Association rules extracted more general relationships
- Combined approach had substantially less noise:
  - *97% of relationships were good or very good*

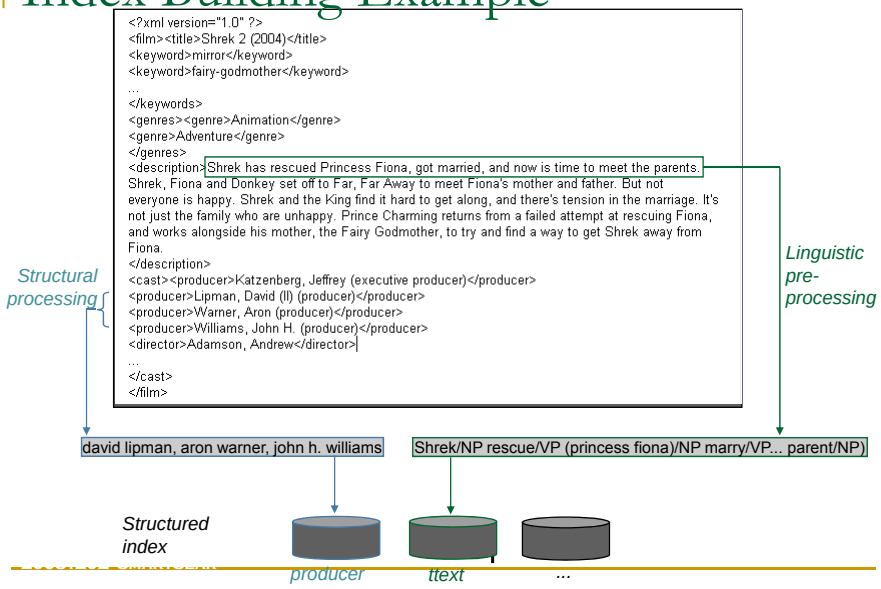


Deutsche Telekom Case

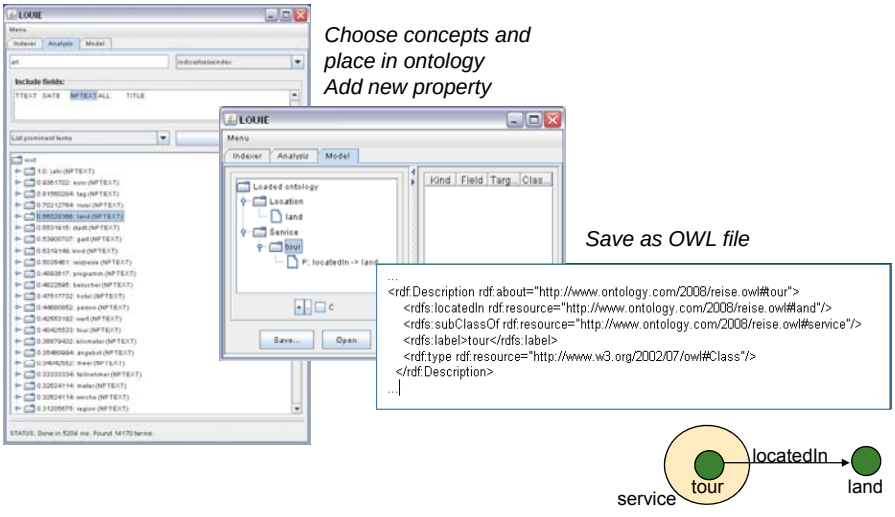
- Workbench developed for DT's movie download service
  - Semi-automatic ontology construction
  - Unstructured or XML-structured docment input



Index Building Example

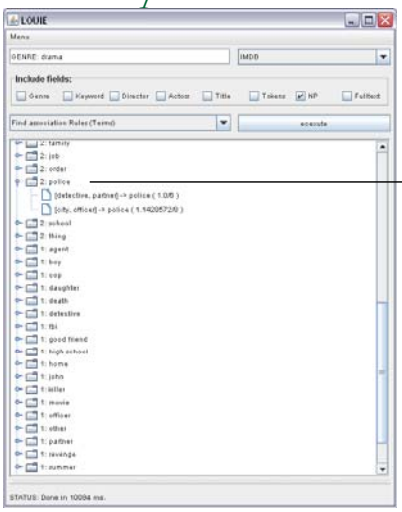


Model Construction Example



Example:

Relationships Relevant to Drama Genre



- Association rules on extracted concepts

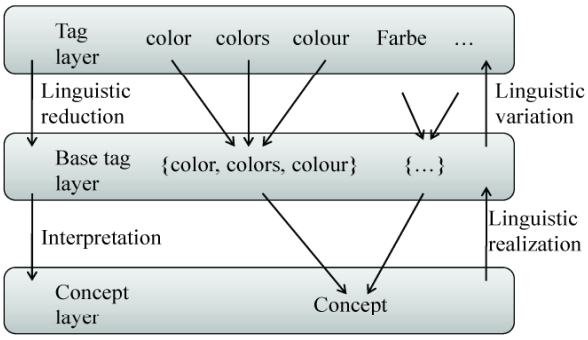
2: police
[detective, partner] -> police (1.0/6)
[city, officer] -> police (1.1428572/8)

What Statistics and Linguistics can Do to Folksonomies

- Continuous evolution of repository
- Uncontrolled production
- Multitude of producers
- Variable quality

- *Semantics useful only if semantic representations aligned with document content*
  - *Concepts*
  - *Relationships*

Are Tags like Terms or Concepts?



- How to identify underlying concept?

Possible Relationships among Tags

Linguistic variation	Description	Examples
Inflectional	Inflections of same lexical word	{horse, horses}
Orthographic	Spelling differences, misspellings, other variations	{color, colour}, {semantic_web, semanticweb, semweb}
Synonyms	Tags with equivalent meaning	{harddisk, harddrive}, {us, usa}
Abstraction	Generalization / specialization, Aggregation	{mail, gmail}, {ui, gui}
Association	Generic semantic relation	{agriculture, permaculture}

From www.delicious.com

Analyzing Tag Similarities

- Tag vector:
  Vector tv_j of terms $tf.idf_{i,j}$, where $tf.idf_{i,j}$ denotes the semantic relatedness weight for each term t_i with respect to tag T_j

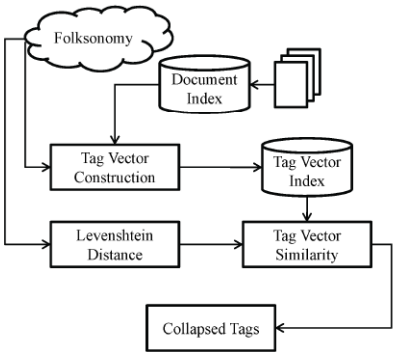
$$tv_{i,j} = \sum_{r \in R} \alpha_{j,r} \cdot f_{i,r}$$

$$tfidf_{i,j} = \frac{tv_{i,j}}{max(tv_{l,j})} \cdot \log \frac{N}{n_i}$$

- $\alpha_{j,r}$ is number of times tag j has been used to tag document r
  $f_{i,r}$ is the frequency of term i in document r.

- Tag vectors contain weighted words that are used to describe the content of the tags

Tag Similarity Score



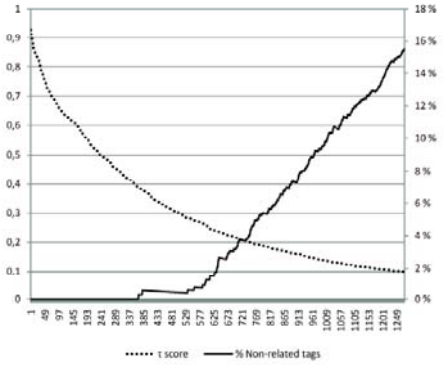
Case study:
:
Crawl of Delicious web site for bookmarks tagged with Wikipedia
228536 bookmarks (195471 used)
51296 users
72420 unique tags
65922 unique URLs

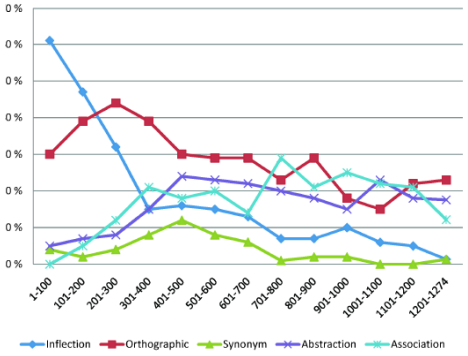
- Combination of Levenshtein edit distance and cosine similarity:

$$\tau_{ij} = (1 - Nlev_{ij}) \cdot cossim(tv_i, tv_j)$$

Some Promising Results

- Low scores mean unrelated
- High scores mean non-semantic relatioships





Taxonomic Tag Relationships

- Method:
  - Construct tag vectors
  - Association rules for tags
  - Cosine similarity for separating types of relationship

```
health
  |--diet
  |--exercise
  |--fitness
  |    |--exercise
  |
  |--nutrition
  |    |--diet
  |
  |--sleep
       |--dreams
       |--polyphasic
```

```
music
  |--audio
  |    |--codec
  |    |--recording
  |    |--sound
  |
  |--hiphop
  |    |--rap
  |
  |--rap
  |--recording
  |--sound
```

- Experimental settings:
  - Delicious training set
  - Results
    - *Synonyms: 6.1% (similarity score 0.78-0.86)*
    - *Correct hierarchical relation: 43.7%*
    - *Different hierarchical relation: 31.7%*
    - *Not related: 5.3% (similarity score 0.02-0.05)*
    - *Unknown: 13.2%*

Conclusions

- Traditional content management technologies suffer from a terminology problem
- Ontologies and Folksonomies may be useful for managing unstructured document content
- Obvious challenges:
  - Ontologies too expensive and time-consuming to build/maintain
  - Tags of variable quality
  - Tags lack sufficient structure
- Statistics & linguistics may be useful…
  - Same techniques useful for both ontology-driven and tag-driven content management