NTNU

Norwegian University of Science and Technology

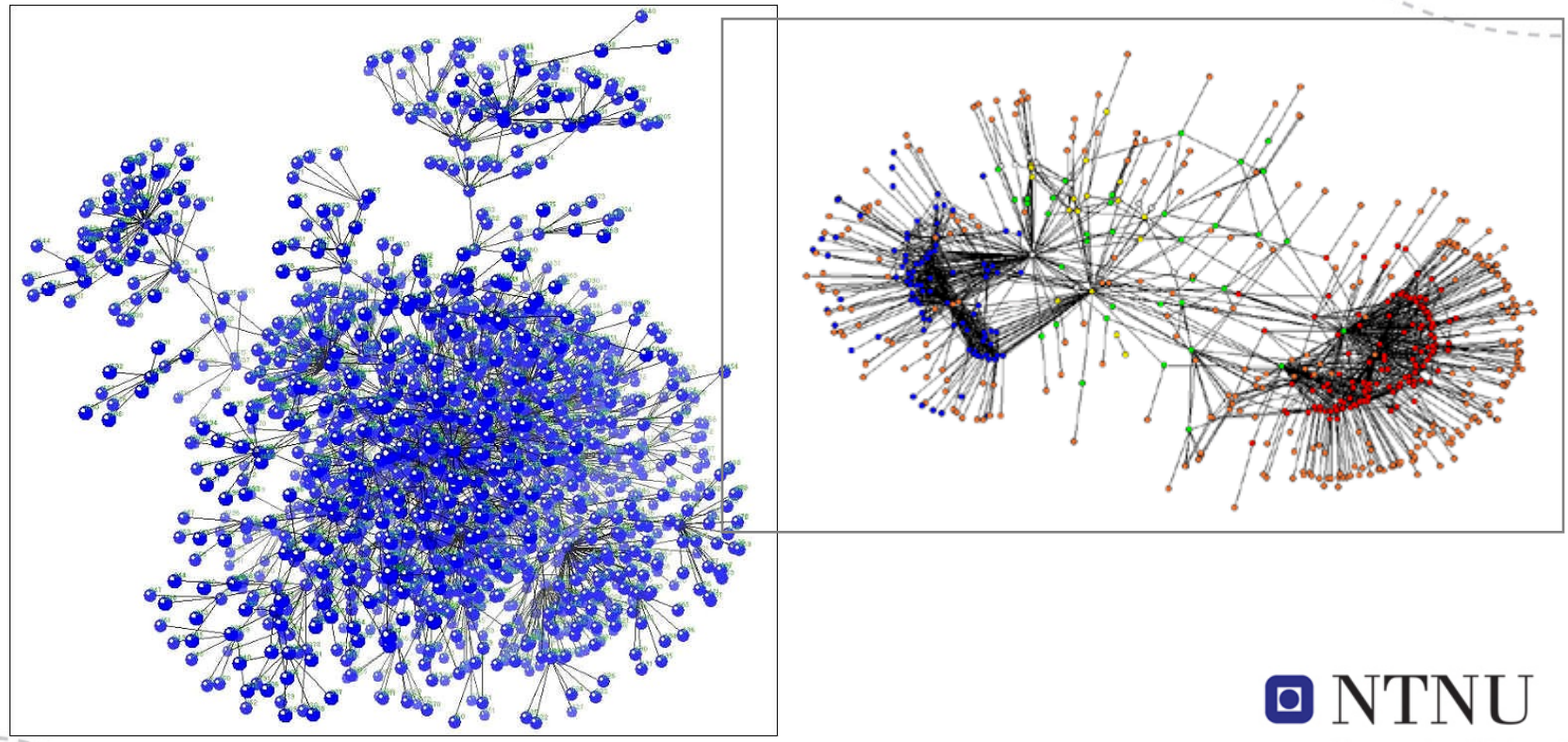
Detecting Semantic Drift in Ontologies

Jon Atle Gulla

Norwegian University of Science and Technology, Trondheim

Ontology Evolution

- How to maintain/assess complex ontologies?



NTNU
Norwegian University of Science and Technology

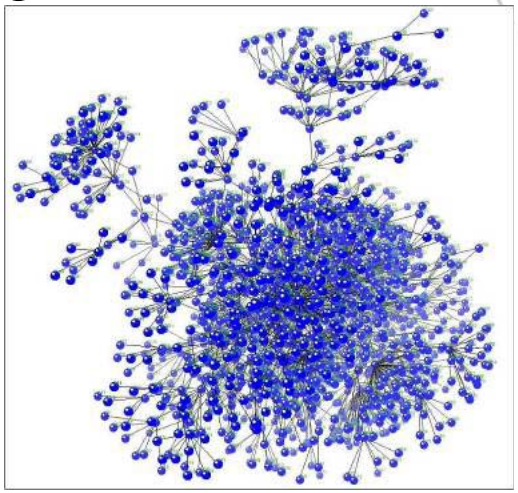
Agenda

- Ontology evolution
- Concept signatures
  - What are signatures?
  - How to construct signatures?
- Quality and Concept Signatures
- Semantic Drift
  - Strength of properties over time
  - Strength of hierarchy over time

Ontology Evolution Issues

- New concepts emerge, others disappear
  - *Incremental ontology learning*
- Hierarchical structures change
- Relationships between concepts change
- Real use of concepts differ from defined meaning

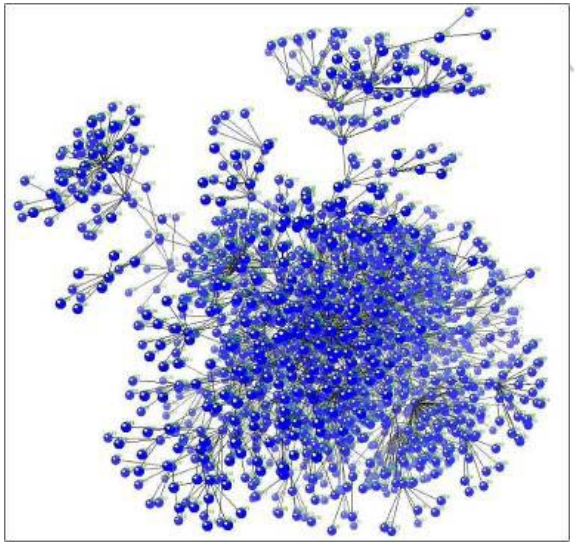
- *Last three addressed by our signature approach to ontology evolution*



NTNU
Norwegian University of Science and Technology

Vector Approach to Ontology Evolution

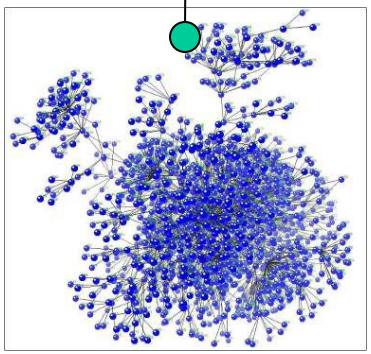
- Concept Signature:
  - Describe the meaning of every concept by means of words that to some extent are related to the concept
    - Abbreviations
    - Synonyms
    - Related words
    - Generalized or specialized terms
    - Etc.
  - Extract descriptions automatically from ontologically categorized text
  - Define quality properties of ontologies in terms of properties of these descriptions:
    - Similarities
    - Subsumption



NTNU
Norwegian University of Science and Technology

Concept Signatures

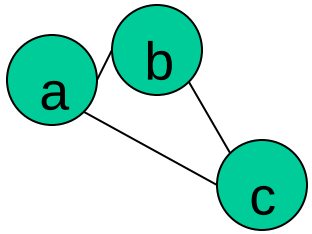
CAR →



References	*Synonyms*	*Concept signature*
Car, bil, voiture, auto,…	Car, automobile, motor vehicle,	1.0: Car, 1.0: automobile, 0.9:motor vehicle, 0.7: SUV, 0.7: cabriolet, 0.2: vw, 0.01: driver …

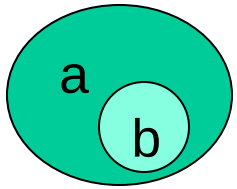
Signature Analysis

- Similarity:



- A & b more similar than a & c
  - A and b share many terms
  - A and b describe some of the same reality
  - A and b semantically close
  - A and b's concepts should be semantically related in the ontology

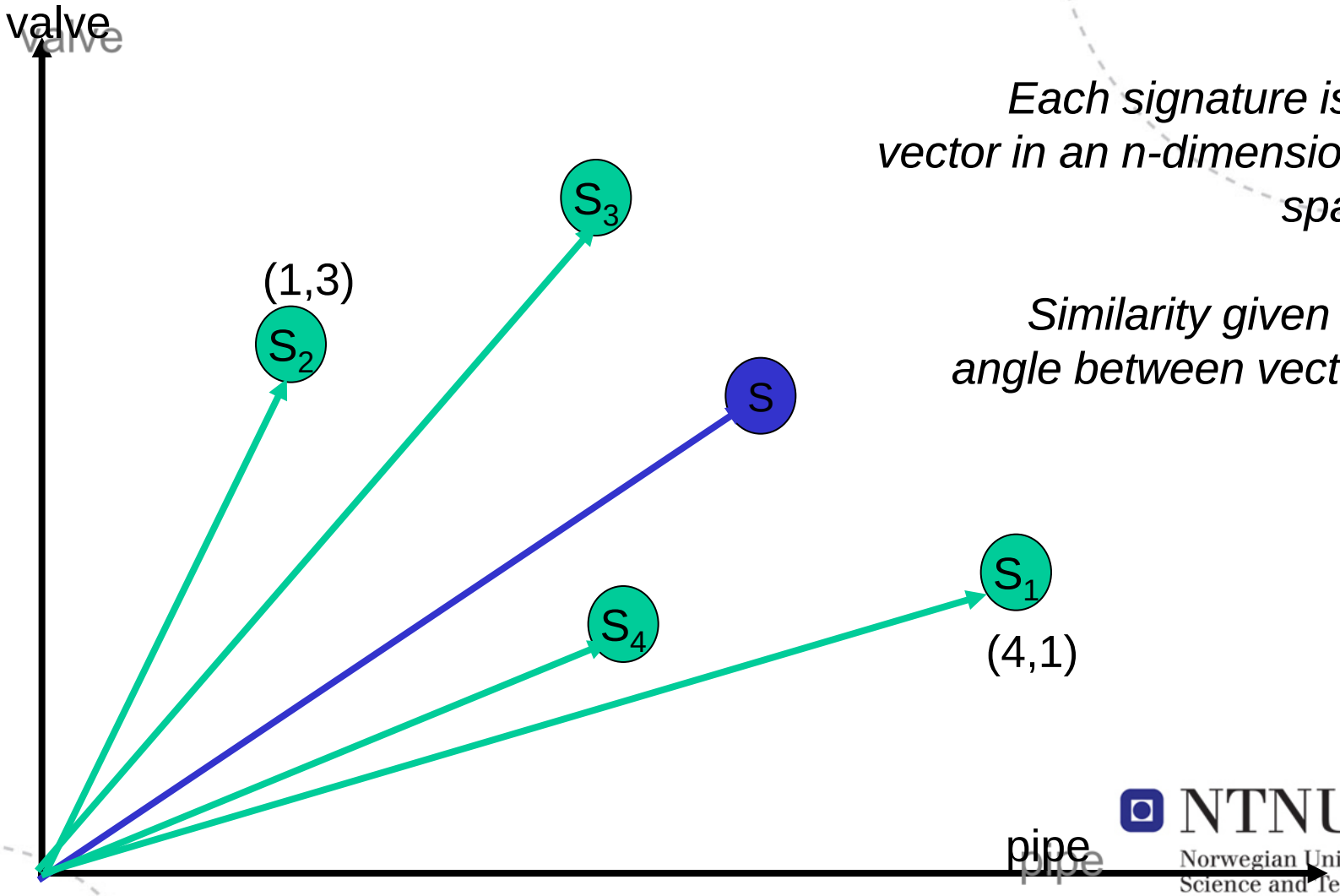
- Subsumption:



- B is contained in a.
  - A describe the same reality as b, plus more
  - Taxonomic relationship
    - A has more abstract descriptions
    - A has more specialized descriptions

NTNU
Norwegian University of Science and Technology

Similarity of Concept Signatures



Each signature is a vector in an n-dimensional space

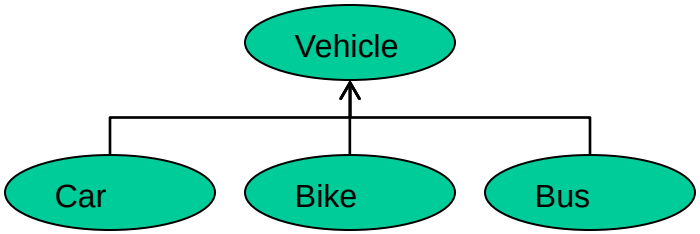
Similarity given by angle between vectors

NTNU
Norwegian University of Science and Technology

What can we do with Concept Signatures?

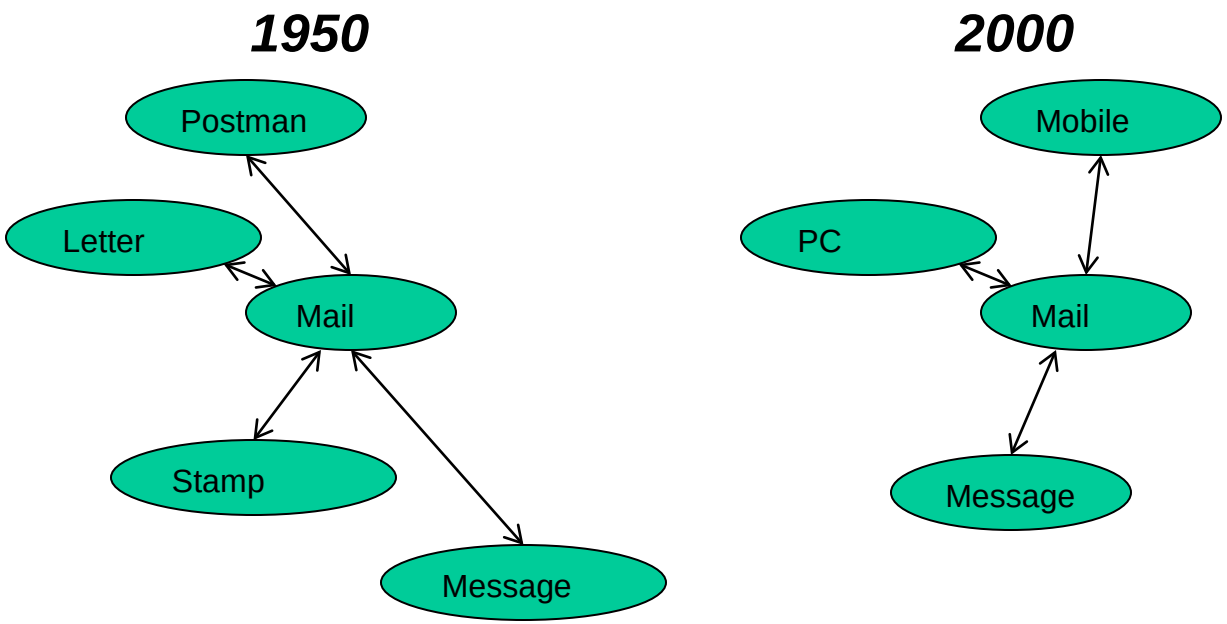
1. Assessing Quality from Signatures

- 1. The relationship between super and sub class is stronger than between the sub classes.
- 2. Characterizations of super class and sub class overlap semantically, but refer to different levels of abstraction
- 3. Commonalities among subclasses are defined by their super class.
- 4. There are abstract features of a superclass that are not shared by any subclass.



2. Detecting Semantic Drift

- Concept similarities over time show how concepts's meaning change with respect to other concepts



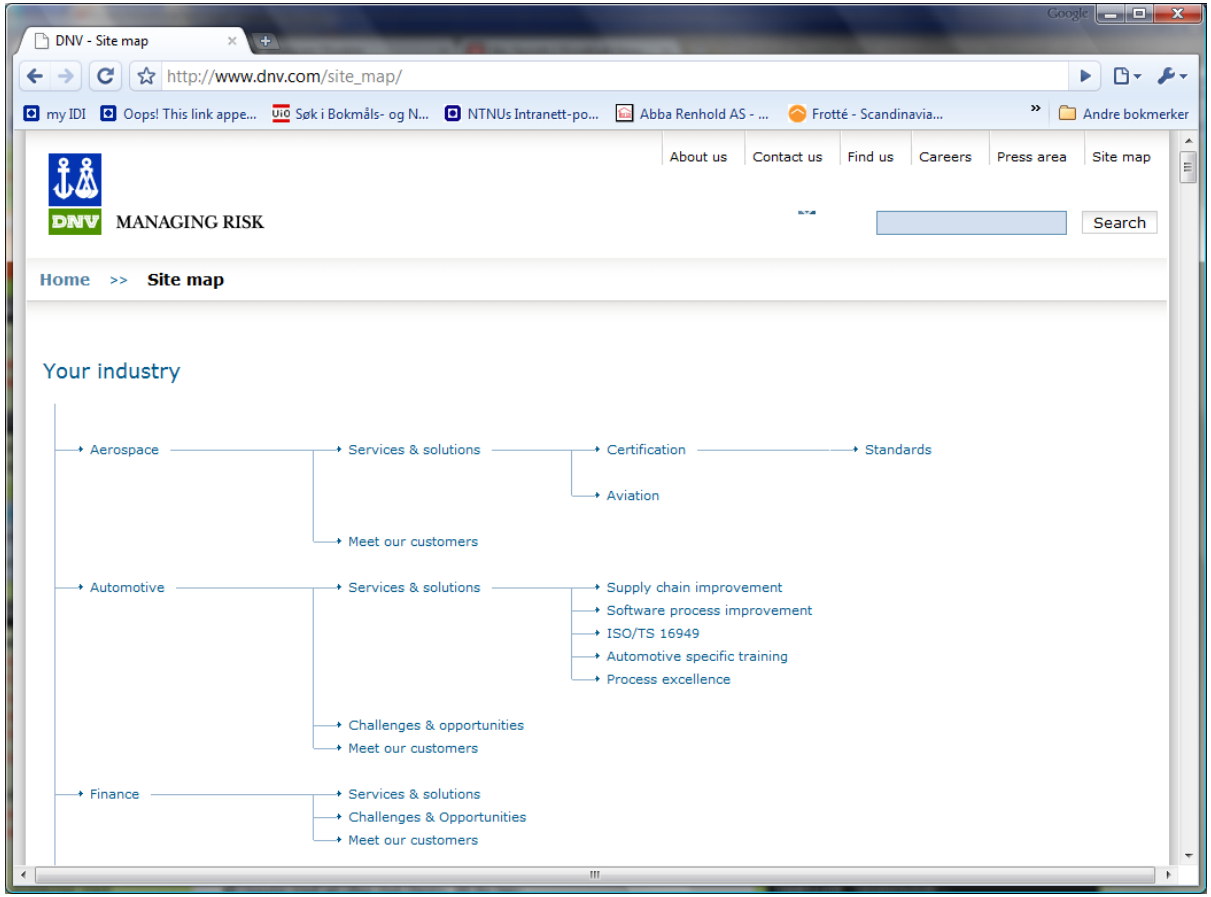
Experimental Setup

- DNV's homepage as ontology
  - Every page defines a concept
  - Text on page describes content of corresponding concept (concept signatures)
  - Site map constitutes a taxonomic structure
- Analysis based on data from 2004 (227 concepts) and 2008 (369 concepts)

- *Testing 5 hypotheses of taxonomic quality*
- *Analyzing possible drift of concepts*

NTNU
Norwegian University of Science and Technology

DNV's homepage

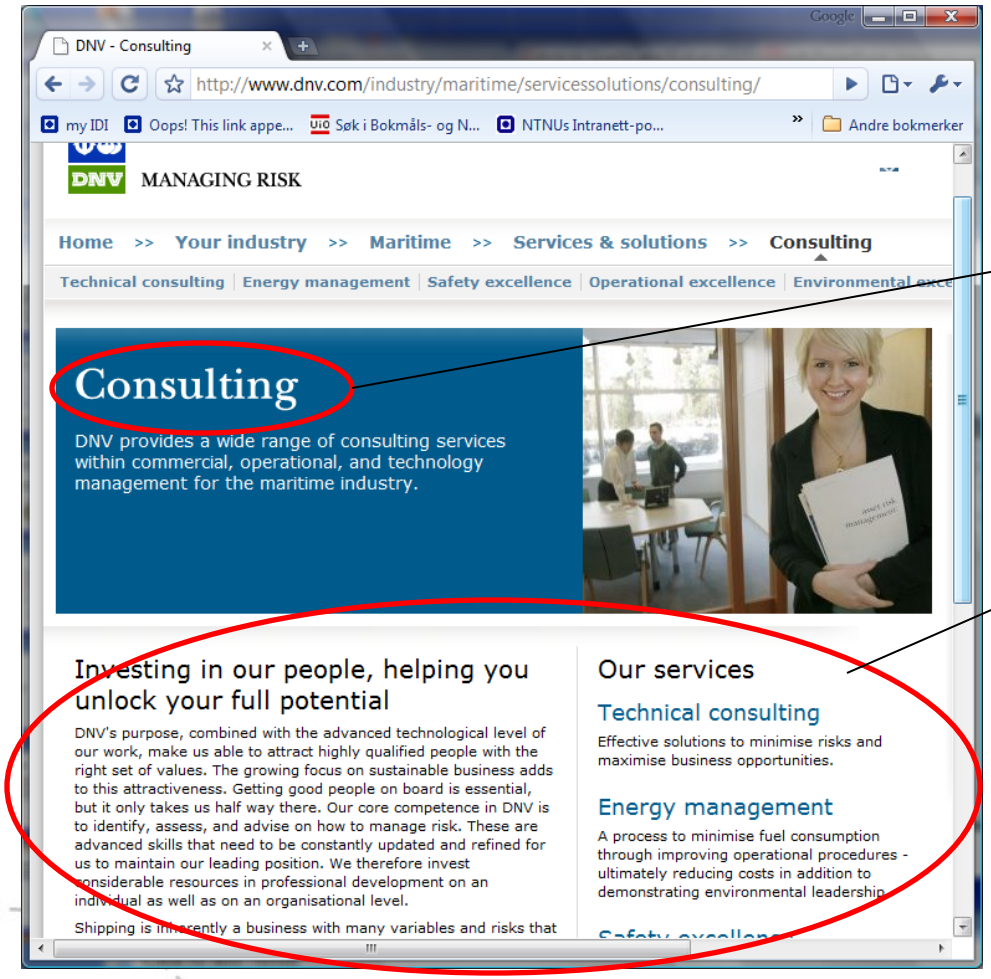


Ontology structure

Each node (page) has a title (concept name) and a text (concept description)

NTNU
Norwegian University of Science and Technology

Understanding Concepts



Concept name

Concept signature source, i.e. our understanding of 'Consulting'

Noun phrases!

NTNU
Norwegian University of Science and Technology

Building Concept Signatures

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)}

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

NTNU
Norwegian University of Science and Technology

’Consulting’ Signatures

2004

Phrasal signature part		Single word signature part
4.63	process industry	
4.63	advanced cross-disciplinary	
2.66	international clients	
2.66	effective risk handling	
2.66	fast-moving world	
2.66	strong business orientation	
2.66	international experience	
2.66	improved health	
2.66	firm base	
2.66	genuine industry knowledge	
2.66	worldwide network	
2.66	strong technological compet	
2.66	enhanced public confidence	
2.66	direct savings	
2.66	unique independence	
2.66	technology competencies	
2.66	better safety management	
2.66	full access	
2.31	experienced consultants	
2.11	environmental performance	

2008

Phrasal signature part		Single word signature part	
5.91	efta inspection	1.227	efta
5.91	real performance	0.567	risk
5.21	industry best practices	0.553	softwar
4.81	risk management services	0.550	consult
4.81	right questions	0.549	knowledg
4.52	business functions	0.506	smart
4.30	operational excellence	0.497	inspect
3.71	knowledge management	0.480	busi
3.20	improvement opportunities	0.475	function
2.95	friday last week	0.424	manag
2.95	ict systems	0.415	abil
2.95	new premises	0.396	object
2.95	norwegian competition authorities	0.376	real
2.95	høvik	0.348	uncertainti
2.95	efta surveillance authority	0.337	question
2.95	efta team	0.326	technolog
2.95	other asset	0.307	complex
2.95	onboard dnv navigator	0.306	â
2.95	management control	0.306	km
2.95	smart ways	0.306	columbia
2.95	telecoms contract	0.306	copyright
2.95	columbia shipmanagement	0.290	improv
2.95	clients she threats	0.283	surveil
2.95	systems functionality	0.280	privaci
2.95	significant risk factor		
2.95	environment risk management		
2.95	in-depth industry insight		
2.95	smart organizations		

NTNU
Norwegian University of Science and Technology

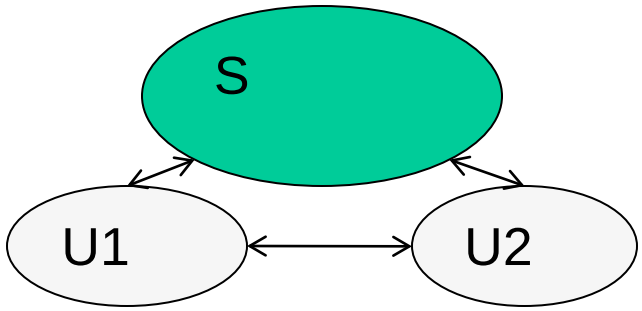
Quality Metric 1

- The relationship between super and sub class is stronger than between the sub classes
- Vector similarity:

$$sim(C_i, C_j) = \frac{\sum_{n=1}^{t} w_{n,i} \times w_{n,j}}{\sqrt{\sum_{n=1}^{t} w_{n,i}^2} \times \sqrt{\sum_{n=1}^{t} w_{n,j}^2}}$$

- Results:

Variable	2004	2008
Mean sub-super similarity	0.347	0.348
Mean sibling similarity	0.197	0.219
Numer of concepts having a mean sibling similarity larger than mean sub-super similarity	5	6



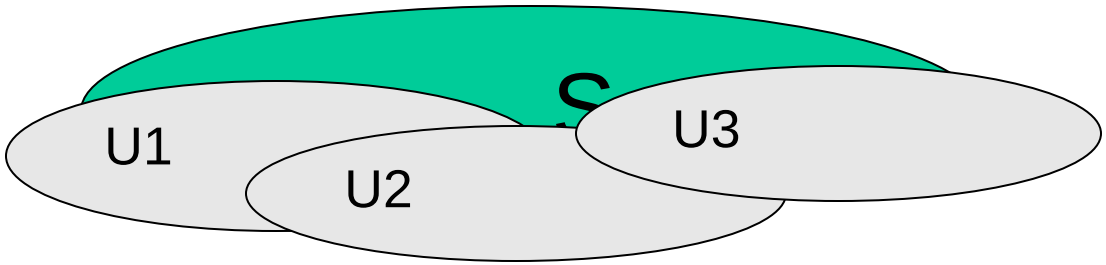
NTNU
Norwegian University of Science and Technology

Quality Metric 2

- Characterizations of super class and sub class overlap semantically, but refer to different levels of abstraction
- Results:

Variable	2004	2008
Mean number of terms in $S^{'}$	65.5	71.3
Mean number of terms in $U_i^{'}$	65.8	71.4
Mean number of terms in $S^{'} \setminus U_i^{'}$	43.3	45.8
Mean number of terms in $U_i^{'} \setminus S^{'}$	41.7	46.6

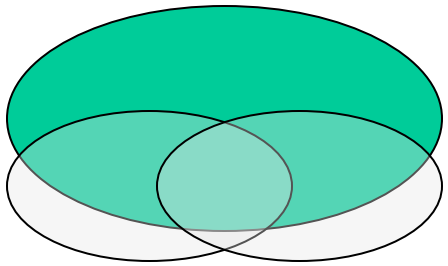
- Interpretation:



Quality Metric 3

- Commonalities among subclasses are defined by their super class.
- Results:

Variable	2004	2008
Mean number of terms in S'	65.5	71.3
Mean number of terms in $(\bigcap_{i=1}^{n} U_i')$	13.7	18.4
Mean number of terms in $(\bigcap_{i=1}^{n} U_i') \setminus S'$	0.7	3.7
Empty result sets	18	28

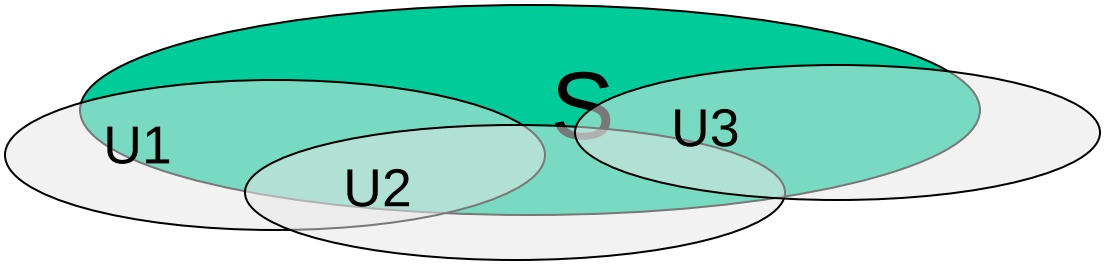


Quality Metric 4

- There are abstract features of a superclass that are not shared by any subclass.
- Results:

Variable	2004	2008
Mean number of terms in $S^{'}$	65.5	71.3
Mean number of terms in $(\bigcup_{i=1}^{n} U_i^{'})$	222.2	223.7
Mean number of terms in $S^{'} \setminus (\bigcup_{i=1}^{n} U_i^{'})$	15.3	23.0
Empty result sets	1	1

- Interpretation:



Detecting Semantic Drift

- Generate concept signatures for all concepts for time t1 and t2
  - Express our understanding of the concepts at t1 and t2
  - Detect small changes of meaning over time
- Calculate similarities between all concepts for time t1 and t2
  - Express new non-taxonomic relationships among concepts
  - Detect changes of existing non-taxonomic or taxonomic relationships

Consulting's Real Relations in 2004

- Ranked list of related concepts reflects people's use of concepts, not how they are currently modeled

0.313	process_industry\process
0.233	asset_operation\asset operations
0.227	maritime\seaskill\competencemanagementcertification\competence management certification
0.225	oil_gas\oil og gas
0.213	maritime\seaskill\seaskill
0.199	consulting\process\process
0.198	technologyservices\whydnv\why dnv
0.186	maritime\maritimeconsulting\maritime consulting
0.181	maritime\seaskill\qsm\quality, structure and measuralble results
0.181	consulting\otherindustries\other industries
0.173	maritime\seaskill\certificationstandards\standards og certificates
0.172	publications\oilgas_news\oil og gas news
0.166	technologyservices\trainee\trainee programme
0.160	consulting\safetyhealthenvironment\managementsystems\management systems
0.158	certificaion\managementsystems\healthandsafety\health and safety

Consulting's Real Relations in 2008

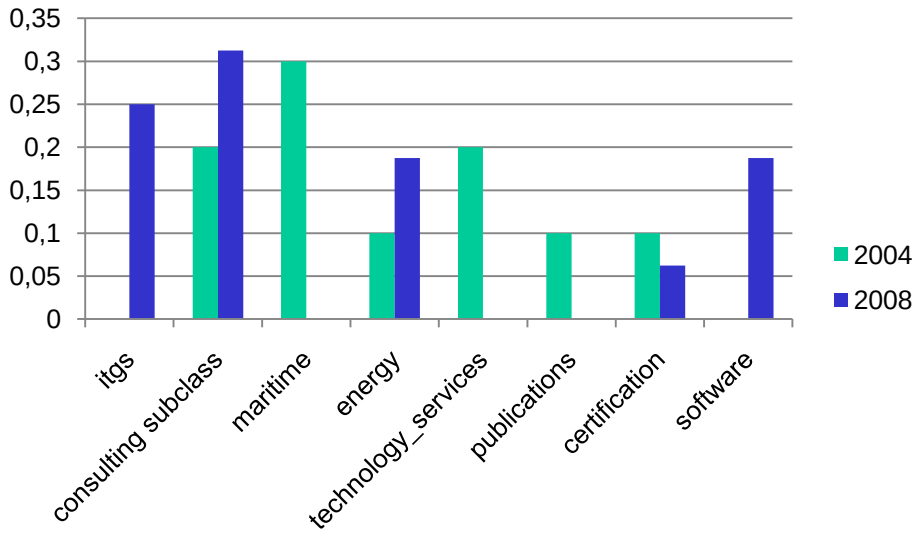
- Consulting closer related to other sectors…

0.447	itgs\it global services
0.258	itgs\key_differentiators\key differentiators
0.240	itgs\swsys\software and systems process improvement
0.237	certification\managementsystems\management systems
0.237	consulting\enterpriseriskmanagement\enterprise risk management
0.228	energy\enterpriseriskmanagement\enterprise risk management
0.225	consulting\enterpriseriskmanagement\cmpi\change management and process improvement
0.222	consulting\enterpriseriskmanagement\cwrm\company-wide risk management
0.221	consulting\safetyhealthenvironment\safety, health and environmental risk management
0.217	software\riskassessment\risk assessment
0.216	consulting\generalindustries\aerospace\aerospace
0.212	software\auditmanagment\audit management
0.211	software\dnv software
0.211	consulting\process\process
0.207	energy\energy
0.205	energy\enterpriseriskmanagement\cmpi\change management and process improvement
0.205	publications\annual_reports\ar_2005\thisis_dnv\this is dnv
0.204	software\process\process
0.203	consulting\generalindustries\utilities\utilities
0.199	energy\enterpriseriskmanagement\cwrm\company-wide risk management
0.195	ict\it and telecom
0.193	itgs\info_man\information management

NTNU
Norwegian University of Science and Technology

Cosine Similarity Between Concepts

- What is consulting related to?



- *Should old relationships still be present in the ontology?*
- *What to do with relationships not present in ontology?*

NTNU
Norwegian University of Science and Technology

Consulting from 2004 to 2008

- Less about maritime, certification, etc.
- More related to energy, IT, ICT and software
- More specializations of consulting

- *Are these changes reflected in ontology?*
- *Should these changes be reflected in ontology?*

Strength of Specialization Changes

- What are the important aspects of 'sea skill'?
- Calculate similarities between sea skill and its specializations:

Sea skill	
2004	2008
qsm competence management certification personnel certification training certification - certified courses certification standards	personnel certification - assessment - certification about_seaskill simulator certification training certification competence management certification downloads standard certificates

- *Should qsm and certification standards still be specializations of sea skill in ontology?*

Conclusions

- Concept (class, individual) signatures:
  - Require textual descriptions of concepts
  - Vector of most important noun phrases describing concept
  - Express our everyday interpretation of concept at time t
- Applications:
  - Assess quality of hierarchical structures
  - Detect development of relationships (hierarchies, propeties) over time
  - Part of larger ontology evolution tool set