Data Mining, Ontologies and the Semantic Web

Konstantin Todorov

University of Montpellier 2

April 2013

Outline

Outline

Outline

The Semantic Web

Towards a more intelligent Web

- The Web contains a large volume of data
  - Text, multimedia, maps, locations...
- but: are these data exploited in the best possible way?

- Towards a more intelligent Web
  - Offer new and better services
  - Search and retrieve information in a more efficient manner

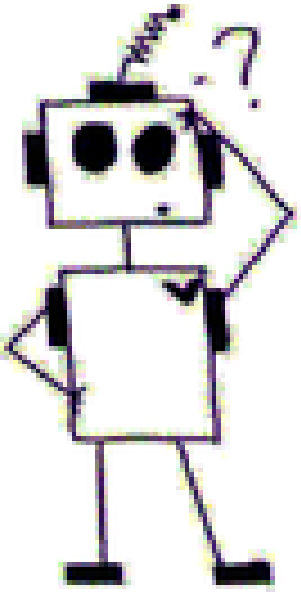
  - Turn data into knowledge

The Semantic Web

The Web of today

- Data on the Web is being created by and for humans
- It is therefore dominated by unstructured or sami-structured documents (text, images, videos, charts,...), linked to one another, and...
- ...comprehensible for humans, but not for machines

- *What do these humans mean?*



The Semantic Web

Pull computers out of their dark age, make them understand semantics.

From Scientific American, 2001.

The Semantic Web

An extension of the Web

"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

Tim Berners-Lee

Formalization and Standardization

- A Web, whose content can be understood and explored by machines
- Completing the informal and unstructured content of the Web of today by formal knowledge
- Define formal languages to describe, explore and reason over the content of the web resources
- Different degrees of formalization will be able to co-exist

Integration

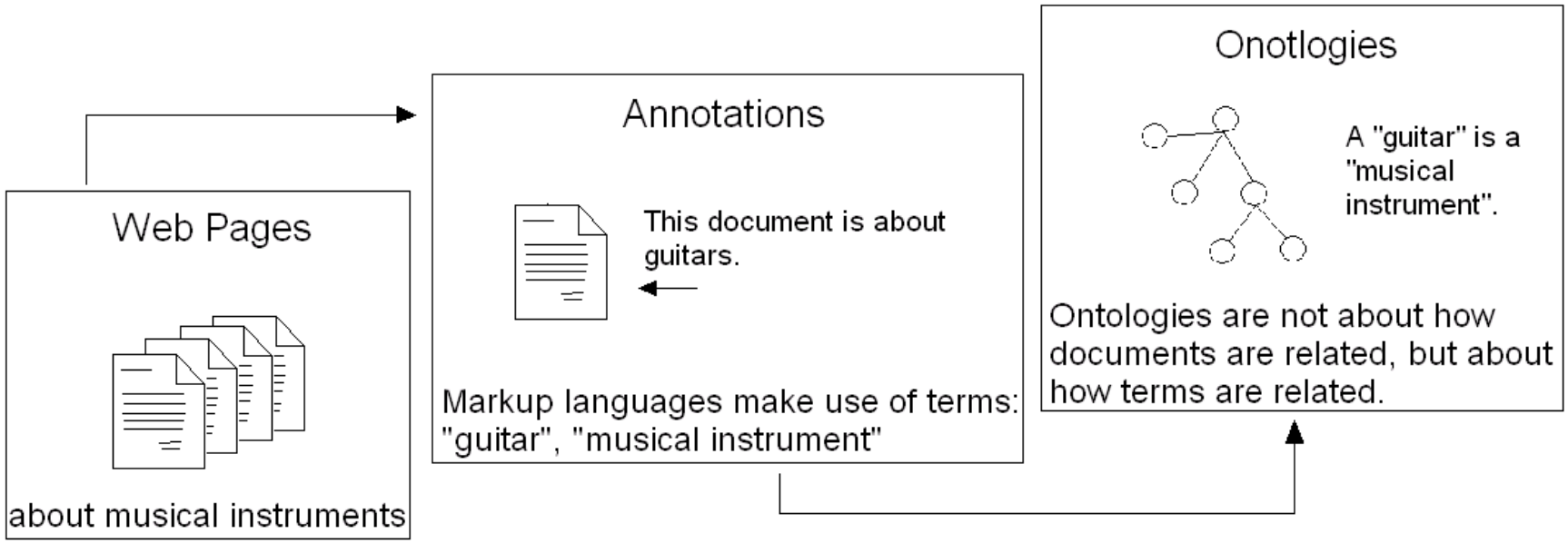
- Integration of heterogeneous data, information and resources
- Automatic combination of web services
- –> Bring information retrieval to a new level

Ontologies

- Knowledge is described by ontologies

The Semantic Web

The role of ontologies



Unstructured data Structure Semantics

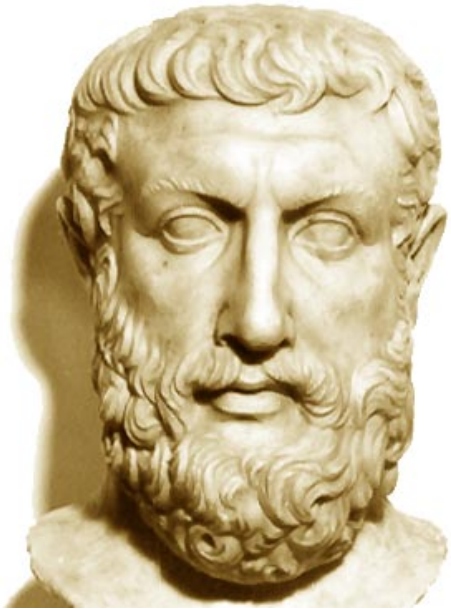
Outline

Ontologies

In philosophy

- The study of what there is, of what exists
- A characterization of the fundamental nature of existence

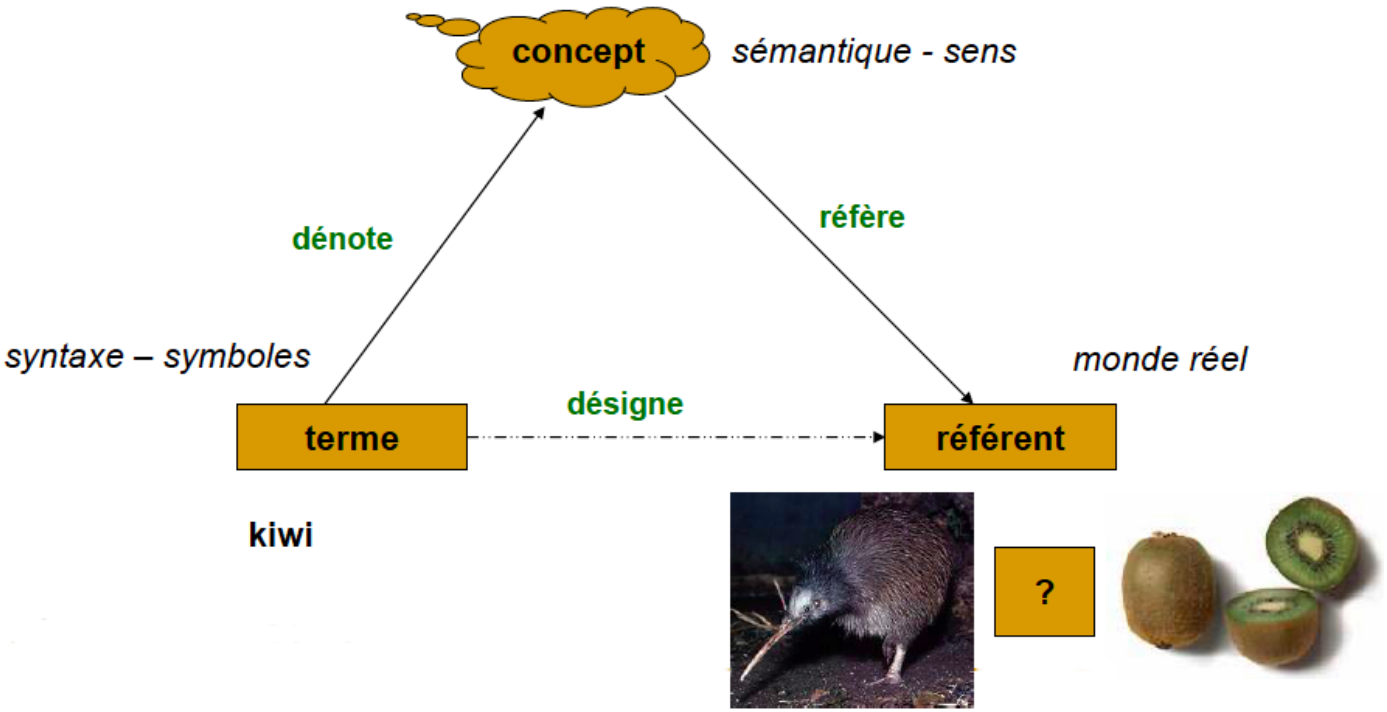
Parmenides, 5th cent. BCE



Parmenides

Ontologies

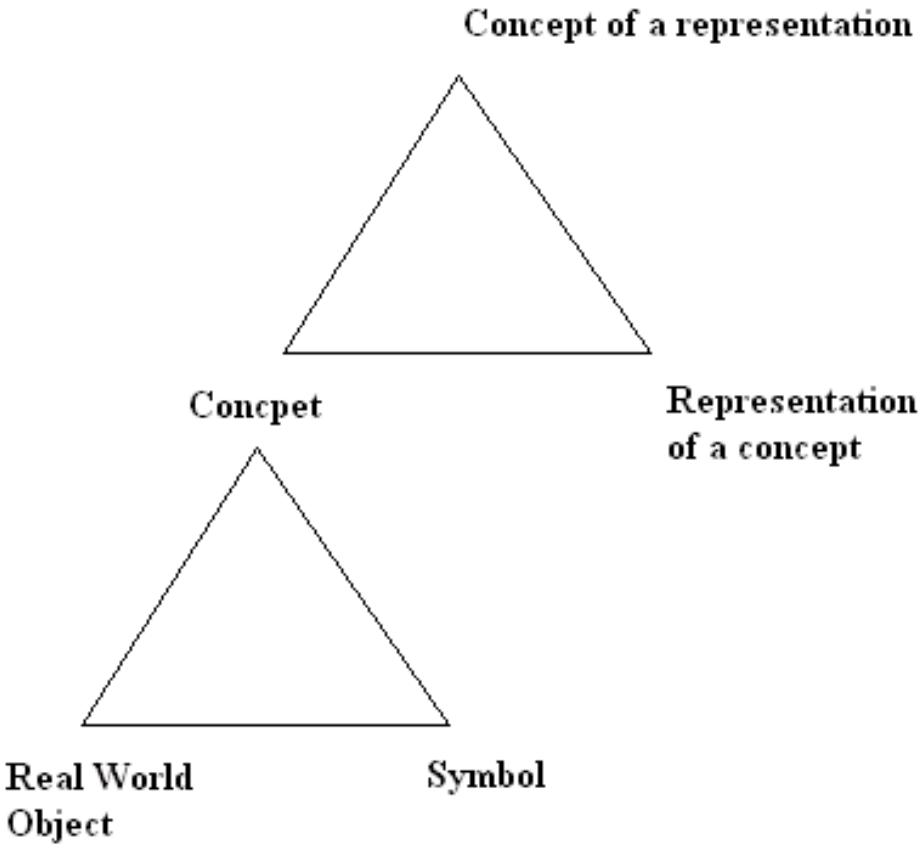
Ogden and Richards (1936) introduced the so called *meaning triangle*:



A **symbol** stands for a **real world object** and evokes a **concept**.
A **concept** refers to a **real world object** designated by a **symbol**.

Ontologies

On top of it, Sowa (2000) built the *knowledge representation triangle*:



How a concept connects to a certain conceptual representation.

Ontologies

Then, what is an ontology?

An ontology is

- a model of some *aspect* of the world
- an *explicit* description of a domain of interest
- a *common vocabulary* for a *shared understanding*
- a specification of the *semantics* of terms
  - Ex: A *student* is a *person* who studies at some *University*
- about *concepts* and how they are related
- *formalized* using a suitable logic

=> Many definitions...

Ontologies

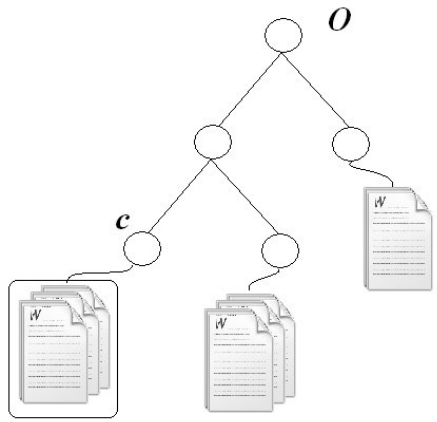
Gruber's definition (1993)

"A **formal specification** of a **shared conceptualization** of a **domain** of interest."

- Formal specification: given in a formal language, thus executable
- Shared: regards a group of persons who agree on a given representation
- Conceptualization: it is about the concepts and how they relate to each other
- Domain: somewhere on the scale "application-driven – universally true" ("concrete – abstract")

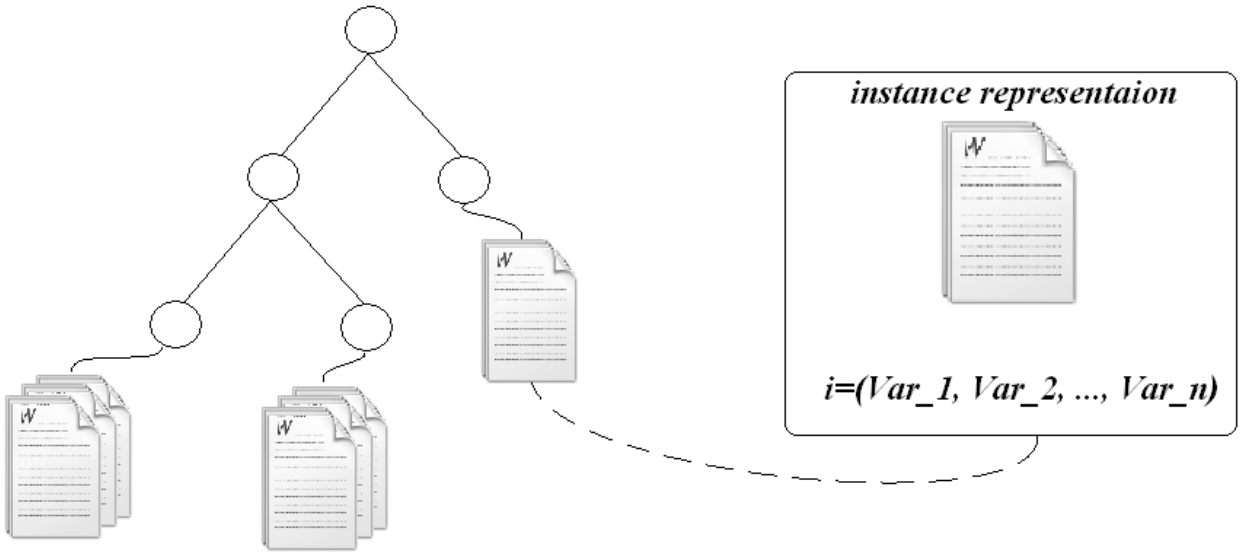
A populated ontology

- $O = \{C, is\_a, R, I, g\}$
- C is a set whose elements are called **concepts**
- *is_a* is a **partial order** on C
- R is a set of other (binary) **relations** holding between the concepts from the set C
- I is a set whose elements are called **instances**
- $g : C \rightarrow 2^I$ is an injection from the set of concepts to the set of subsets of I



Ontologies

Concept instances



The instances:

- define a concept extensionally
- can be *text documents*, *images*, *objects identified by URIs*,...
- can be represented as (real-valued) vectors defined by a set of input *variables* of some kind (the same for all instances in *I*)

Ontologies

Examples

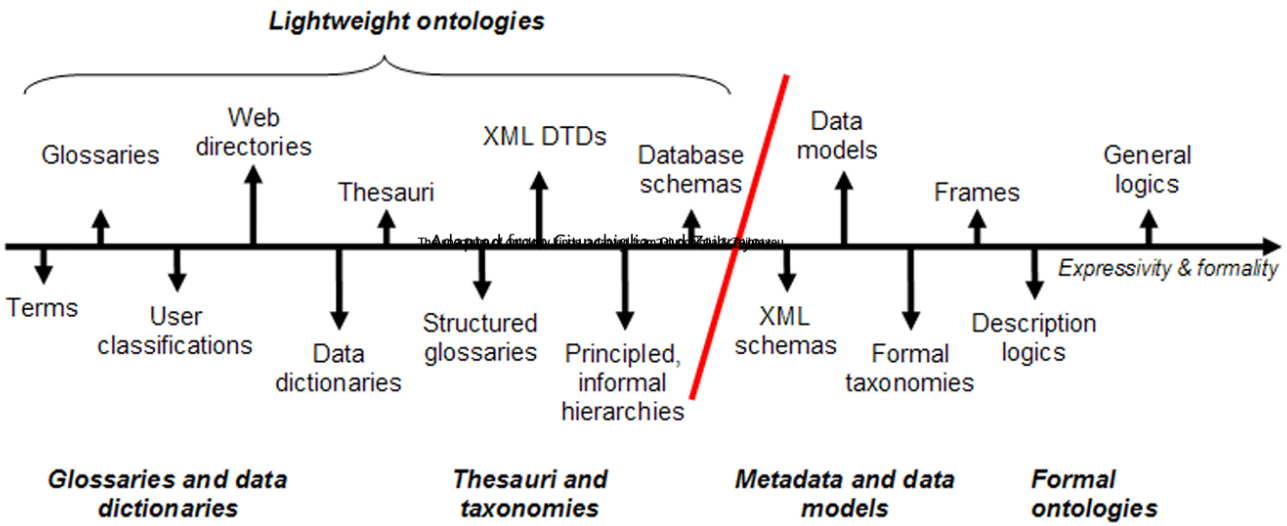
- Web taxonomies
  - Yahoo categories
- Online catalogues
  - Amazon
- Domain specific terminologies
  - FMA - medical ontology

Ontologies

Types of ontologies

Different levels of abstraction and of detail, different application purposes.

- Application ontologies, Domain ontologies, Core ontologies, Top ontologies



–> Expressiveness –>

Ontologies

Applications I

- Ontologies as a specification of a common vocabulary
  - Knowledge sharing
  - Knowledge re-use
  - Collaborativity
  - Assist the development of information systems
- Ontologies for mutual understanding
  - Communication between humans
  - Understanding between humans and software agents
    - support of the core ideas of the SW, web resources search and use
  - Communication between software agents

Ontologies

Applications II

- Ontologies for data sharing
  - Data heterogeneity
  - Data Integration
- Ontologies for information retrieval
  - A vocabulary for annotation of web resources
  - Use hierarchy and class relations in order to interpret this vocabulary
  - Access large collections of data (text, multimedia)
  - Assist user query formulation
  - Query expansion, reformulation

Data Mining and Ontologies?

What can data mining do for ontologies?

On the Web-scale, data mining is applied for

- Web content mining
- Web structure mining
- Web usage mining

Data mining techniques are used to

- learn ontologies
- match ontologies
- provide semantic annotations

Data Mining and Ontologies?

And the other way round?

What can ontologies do for data mining?

Data Mining and Ontologies?

And the other way round?

What can ontologies do for data mining?

– Describe and exchange data for use of ML techniques
– Provide important heuristics in the form of background or domain knowledge to support ML
– Information retrieval supported by ontologies
– Help understand the results obtained from data mining

Outline

- Abstract:
  - What knowledge?
  - What perspective?
  - What application scope?
  - What degree of detail, granularity?
- Represent:
  - What formalism?
  - How to represent the abstraction in that formalism?

- Where to start from?
  - From nothing, from text corpora, from web resources, from existing ontologies...
- Manually or automatically?
  - Different degrees of user involvement

- How to identify the concepts that we are going to use?
- How to find these concepts and where?
- Which among them to keep?
- How to define them?
- How to define their relations, properties?
- How to group them together, how to structure them?

- Construct
  - From human resources, from text or multimedia data, from databases
- Validate
  - Verify the coherence of the resulting ontology
  - Experts validation
  - User validation

Outline

Ontology Learning from Text

A definition (W. Wong, 2009)

The process of identifying terms, concepts, relations and optionally, axioms from natural language text, and using them to construct and maintain an ontology.

An interdisciplinary topic

- **text and data mining, machine learning:** extract rules and patterns out of massive datasets in a supervised or unsupervised manner based on extensive statistical analysis
- **natural language processing:** analyzing natural language text on various language levels (e.g. morphology, syntax, semantics) to uncover concept representations and relations through linguistic cues
- **information retrieval:** algorithms to analyze associations between concepts in texts using vectors, matrices and probabilistic theorems
- **knowledge representation, reasoning:** enables elements to be formally specified and represented such that new knowledge can be deduced

Ontology Learning from Text

The outputs of ontology learning I

Five types of output:

- terms
- concepts
- taxonomic relations
- non-taxonomic relations
- axioms

Terms

- Lexical realization of all that is important in a domain
- Single words, multi-words
- Tasks:
  - preprocess texts: input text format
  - extract terms: part of speech tagging, sentence parsing

Concepts

- What is a concept? ...that's a tough one!
- For our goals: a group of terms, a class of individuals
- Tasks:
  - form concepts: grouping terms together
  - label concepts: use background knowledge (WordNet?)

Relations

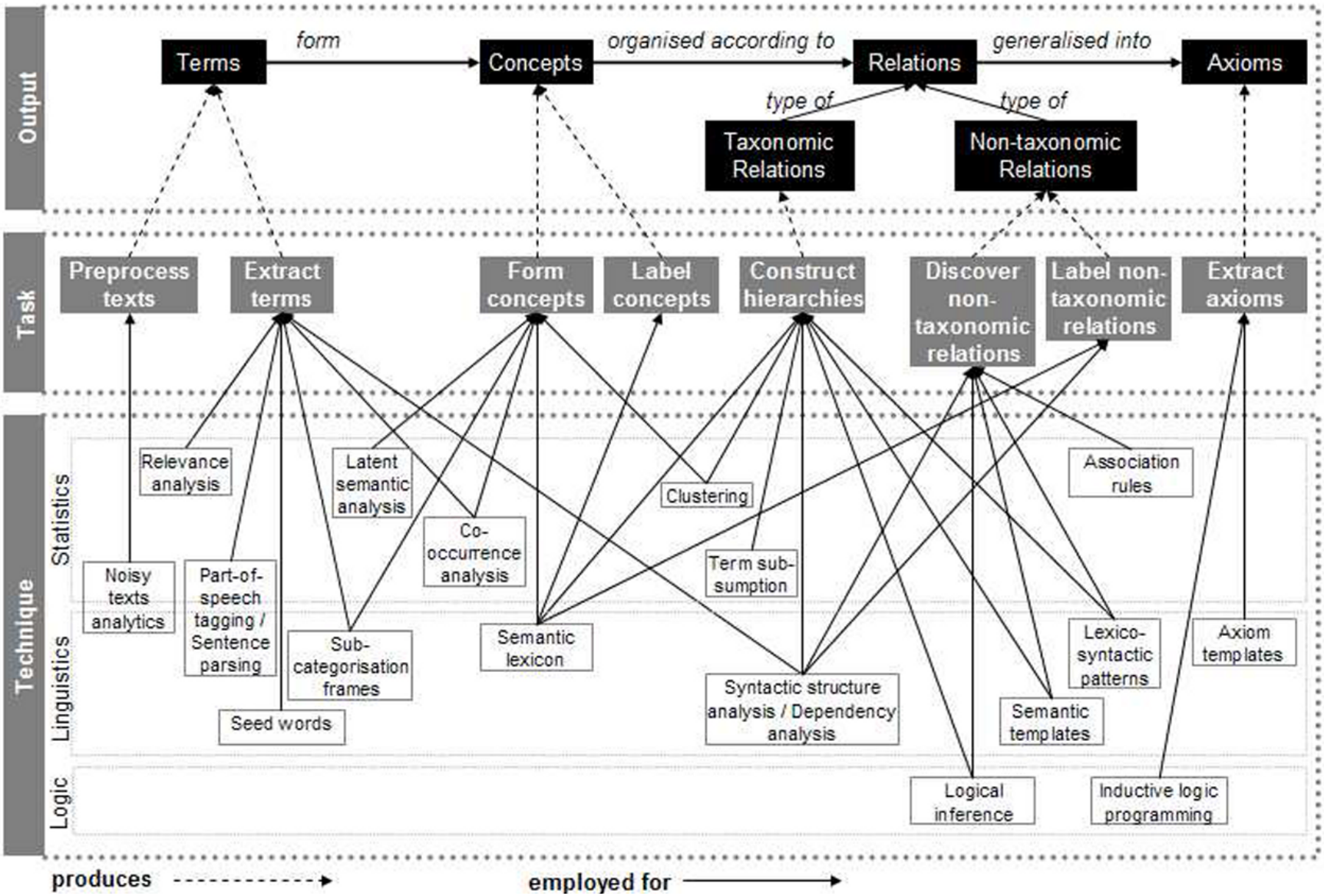
- Structure the concepts
- Hierarchical (taxonomic) or non-hierarchical (non-taxonomic)
- Tasks:
  - construct hierarchy: discovery of hypernyms
    - use of background knowledge, statistical models,...
  - extract non-taxonomic structures: more challenging

Axioms

- Facts that are always taken as true (propositions or sentences)
- Tasks:
  - discover axioms
  - generalization or deduction

Ontology Learning from Text

Outputs, tasks and techniques (W. Wong 2009)



Outline

Techniques for Ontology Learning

Depend on the task to be accomplished

- Statistics-based
- Linguistics-based
- Logic-based
- Hybrid

Fields: Information retrieval, data mining, machine learning

– No consideration of underlying semantics;

– Important at the early stage of ontology acquisition.

- clustering
- latent semantic analysis
- co-occurrence analysis
- term subsumption
- contrastive analysis
- association rule mining

Clustering

– concept formation, taxonomic relations

- Agglomerative
  - Assign terms into groups
  - Using a measure of relatedness
- Divisive
  - Start with all terms and split them into subgroups
- Problem: similarity computation due to high dimension of data
  - => Use of feature-less representation (Normalized Google Distance)

Latent Semantic Analysis

– concept formation

- Dimension reduction techniques
- Reveal inherent "hidden" relations between terms
- Resulting *orthogonal* dimensions:
  $\{(car), (truck), (flower)\} -- > \{(1.3 * car + 0.28 * truck), (flower)\}$
- Problem: complexity

Occurrence and co-occurrence

– term extraction, concept formation

- The presence of two or more terms within a sentence or an N-gram
- Coupled with association strength measures...
  - Mutual Information: measure the discrepancy between the joint probability of two terms and their individual probabilities
    - Estimate on corpora: $PMI(t,s) = log \frac{F_{t,s} \times n}{(F_{t,s}+F_s)(F_{t,s}+F_t)}$
  - Rank Correlations: parameter-free correlation measures, act as similarity measures
- ...or similarity measures (e.g., cosine)

Conditional probabilities

– taxonomic relations

- ...of the occurrence of terms
- Employed to discover hierarchical relations between terms
- Using a term-subsumption measure
  - $P(x|y) > t$ and $P(x|y) > P(y|x)$ for a given threshold t
    - *Example:* $P(''fish''|''shark'') > P(''shark''|''fish'')$
  - Estimate by using corpora:
    - x subsumes y if the documents in which y occurs are a subset of the documents in which x occurs

Relevance analysis

– term extraction

- TF-IDF: term frequency within a document scaled by the inverse document frequency of the term in the corpus
- Evaluate the relevance of a term w.r.t. a document and a collection of documents (the extent of its occurrence in a single document and in a corpus)

Association rule mining

– taxonomic relations, non-taxonomic relations

- associations: $\{chips, beer\}$
- induction: $\{chips, beer\}$, $\{peanuts, soda\}$ -> $\{snacks, drinks\}$

Field: Natural Language Processing

– Applicable at all levels of the ontology learning process

- part-of-speech tagging and sentence parsing
- syntactic structure analysis and dependency analysis
- semantic lexicon, lexico-syntactic patterns, semantic templates, subcategorisation frames, seed words.

Part-of-speech tagging and sentence parsing (syntactic analysis)

– term extraction

- Provide the basis for further linguistic analysis
- Brill Tagger, TreeTagger, GATE, NLTK,...
- Disclaimer: many parsers are actually built on statistical methods and make use of training data in the form of (manually) parsed corpora

Syntactic structure analysis and dependency analysis

– term extraction, taxonomic relations, non-taxonomic relations labeling

- Examine syntactic information to discover terms and relations at a *sentence level*
  - Example: ADJ-NN can be extracted as potential terms, verb-phrases can be ignored
- In dependency analysis, grammatical relations are used (complement, subject, etc...) to discover more complex relations
  - Example: Jane took the book from the library.

Use of a semantic lexicon

– concept formation, concept labeling, relations labeling

- General (WordNet) or domain specific (UMLS[1])
- Access to a large collection of predefined words and relations
- A set of synonyms (sunsets in WordNet) model a concept
- Assigning semantic relations (hyponymy, meronymy)
- Word-sense disambiguation

[1] Unified Medical Language Systems

Lexico-syntactic patterns

– taxonomic relations, non-taxonomic relations

- Extract hypernyms and meronyms
- Use of patterns:

  NP such as NP, NP, ... and NP;

  NP and NP are parts of NP
- Problem: producing such patterns (manually? too costly)

Sub-categorization frames

– term extraction, concept formation, non-taxonomic relations (?)

- Definition: the number and kinds of other words that a word selects when appearing in a sentence
- *Joe wrote a letter.* -> "*write*" selects "*Joe*" and "*letter*" as its subject and object
- Part of the speaker's *knowledge* of the world

Seed words

– term extraction

- Anchors: provide "good" starting points to discover other related terms

Fields: Knowledge Representation, Reasoning, Machine Learning

– Least common in ontology learning;

– Used for discovering relations and axioms.

- Inductive Logic Programming
- Logical Inference

Inductive Logic Programming

– taxonomic and non-taxonomic relations

- A collection of positive and negative examples:

- *"cats have fur", "tigers have fur" -> "felines have fur"*
  positive examples
- *"dogs have fur" -> "mammals have fur"* ?
  positive example
- *"humans do not have fur" -> "canines and felines have fur"*
  negative example

Logical Inference

– axiom discovery

- Derive implicit relations from existing ones
  (using transitivity, inheritance, etc...)
- *"Socrates is a man"*, *"All men are mortal"* –> *"Socrates is mortal"*

- Some transitivity problems:
  "Human eats chicken", *"Chicken eats warms"* –> ?

Outline

Evaluation

According to a perspective

The ontology is not the end result

Rather a *means* to achieve some further goals

How *good* an ontology have we constructed?

Not an easy question.

Good with respect to...

- a given application context
- the "fit" of the ontology to the domain knowledge (in the form of corpora)
- a benchmark
- an expert assessment

Evaluation

According to a layer

Ontologies are complex artifacts, composed by multiple layers.

- **Terminological layer:** correctness of the terminology
  - *Are the terms used to identify a concept included and are they correct?*
- **Conceptual layer:** coverage
  - *How well do the extracted terms cover the domain?*
- **Taxonomical layer:** structure
- **Non-taxonomical layer:** adequacy of the relations

Adopt a **gold-standard** approach:

an *expert ontology* vs. a *learned ontology.*

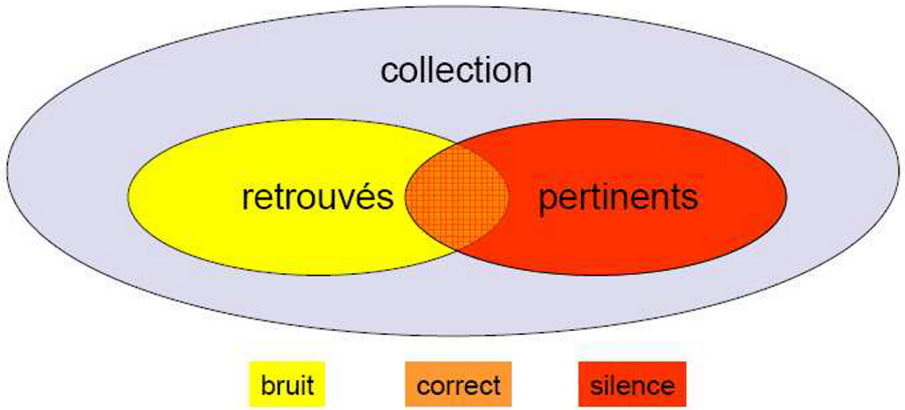
Evaluation

Performance measures at the terminological and conceptual layers I

On the terminological and conceptual layers

- **Precision and recall**

$$P = \frac{relevant\_found}{all\_found}, \qquad R = \frac{relevant\_found}{all\_relevant}.$$



Evaluation

Performance measures at the terminological and conceptual layers II

- **Lexical overlap**

$$LO = \frac{|C_d \cap C_m|}{|C_m|},$$

C_d - discovered concepts, C_m - recommended, $|.|$ - set cardinality. A measure of recall.

- **Ontological Improvement / Ontological Loss**

$$OI = \frac{|C_d - C_m|}{|C_m|}, \qquad OL = \frac{|C_m - C_d|}{|C_m|}.$$

Evaluation

According to a layer

On the structural layers

- **Taxonomical layer:** Structure
- **Non-taxonomical layer:** Adequacy of the relations

These layers are more tough to evaluate!

Types of measures

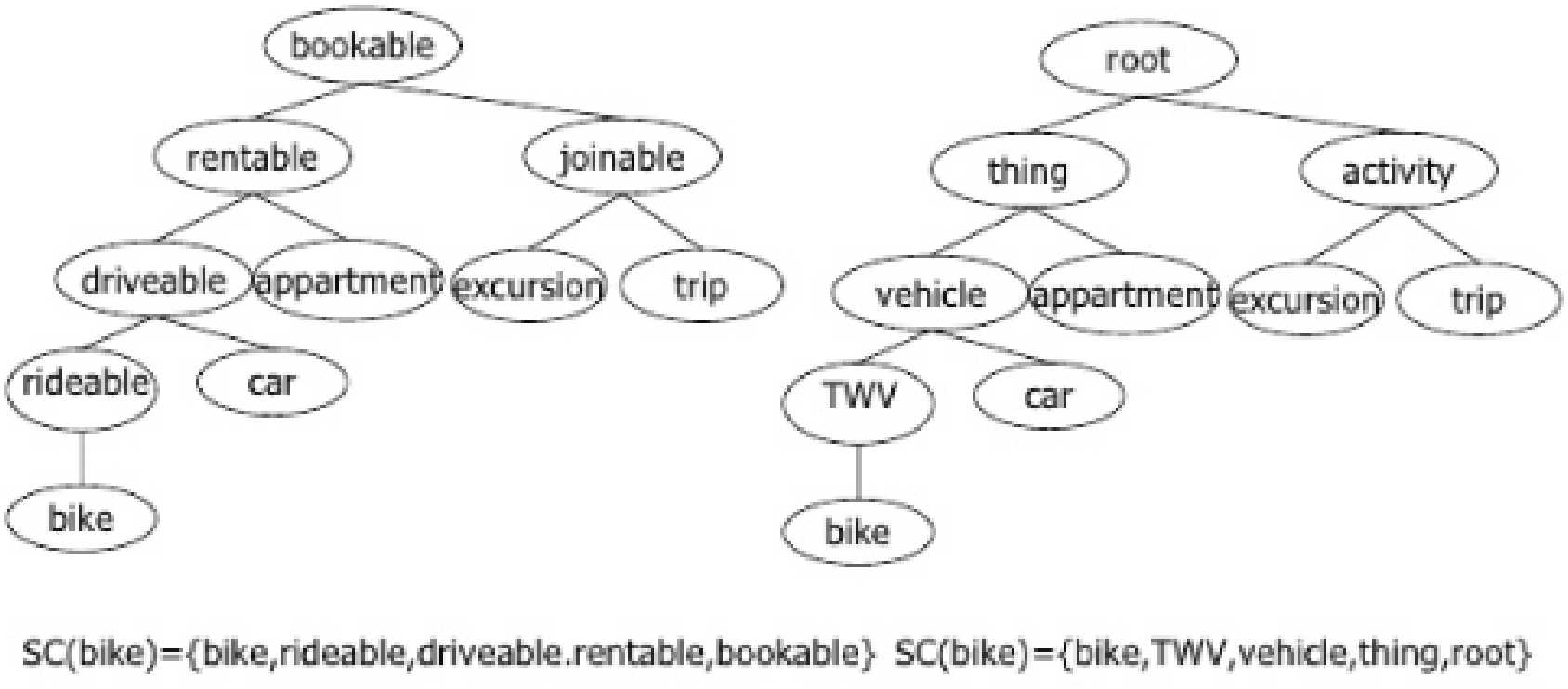
- **Local:** measure the similarity of the concept's position in the learned taxonomy and in the benchmark
- **Global:** average the local scores for all concept pairs

Evaluation

Performance measures at the taxonomical layer II

A common measure: the **taxonomic overlap (TO)**

- Semantic cotopy: the set of all super- and sub-concepts of a given term



Locally: $TO(bike, O_1, O_2) = \{SC_{O_1}(bike)\} \cap \{SC_{O_2}(bike)\} = 1/9$

Evaluation

Performance measures at the taxonomical layer III

From there on, compute the **global** taxonomic overlap:

$$TO_{global}(O_1, O_2) = \frac{1}{|C_1|} \sum_{c \in C_1, c \notin C_2} TO(c, O_1, O_2),$$

where C_1 and C_2 are the sets of concepts of O_1 and O_2, respectively.

Note that TO_{global} *is not* symetric.

A posteriori approach

- Ask a domain expert to evaluate each concept of the learned ontology
- Group concepts in three categories:
  - correct
  - new
  - spurious
- Precision = (correct + new) / (correct + new + spurious)

Several Systems

- OntoLearn, TextToOnto, ASIUM, TextStorm/Clouds, SYNDIKATE,...

Some conclusions:

- Mostly semi-automatic
- Depend on static background knowledge, no flexibility to port on different domains and languages
- A common evaluation platform is still to be provided
- Discovering relations (non-taxonomic) and axioms is still work in progress

Ontology Learning

From multimedia data?...

And what if we have images instead of text documents?...

Ontology Learning

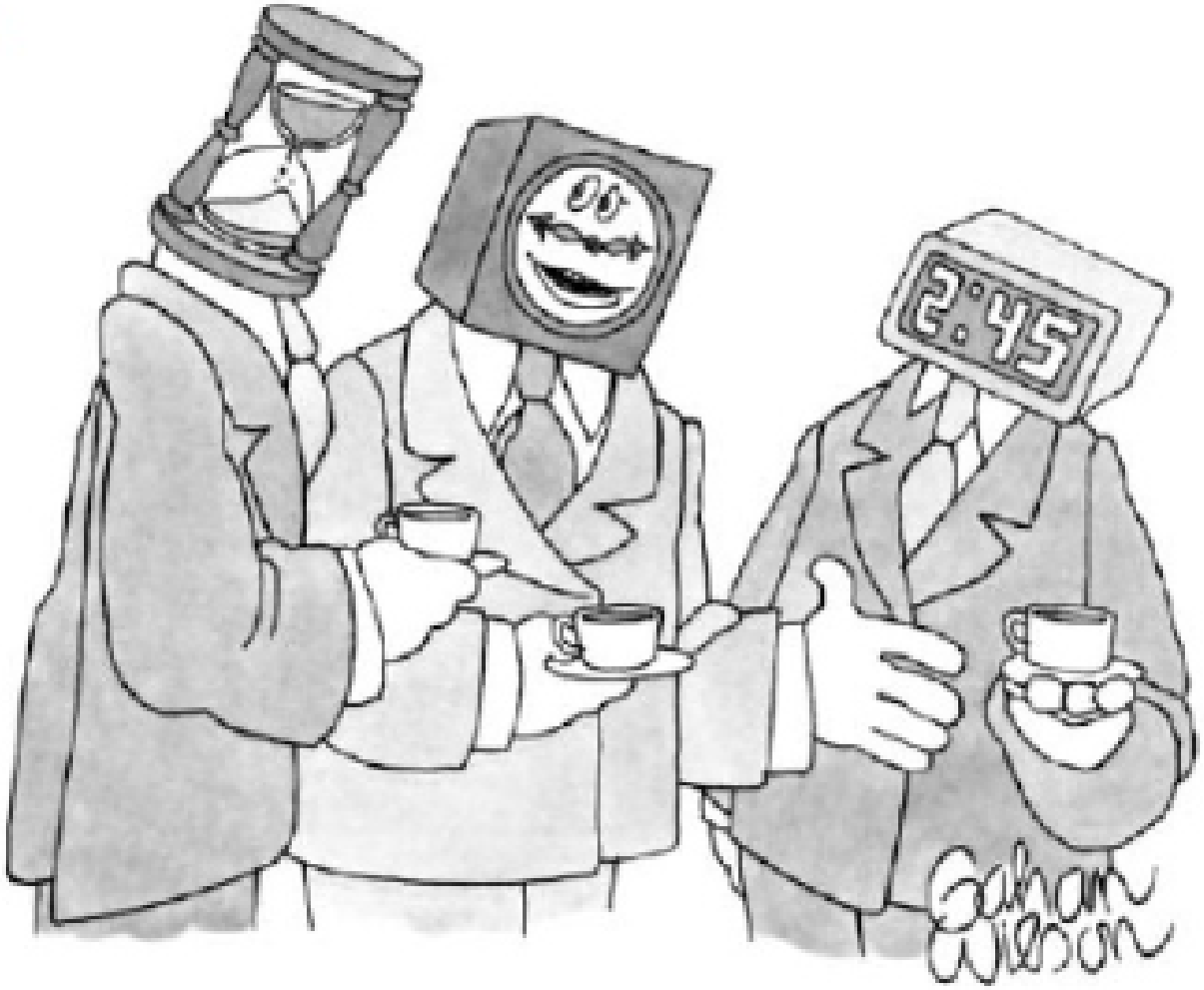
From multimedia data?...

And what if we have images instead of text documents?...

– automatic concept detection in an image (machine learning, classification)
– construction of concept hierarchies from tags and classification methods
– towards a linguistic description of an image, a video
– use of textual information associated to the image?

Outline

Ontology Matching



"*Basically, we're all trying to say the same thing.*"

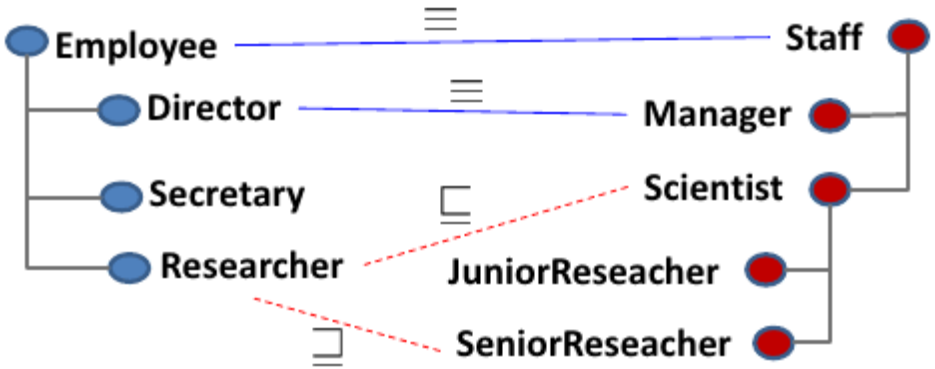
Ontology Matching

Introduction to the problem

Ontologies are created in a **decentralized**, strongly **human biased** manner.
Many ontologies describing the same domain of interest

=> **ontology heterogeneity:**

- syntactic
- terminological
- conceptual / structural



=> **Ontology Matching:** detect the semantic correspondences between the elements of two ontologies.

Ontology Matching

Types of ontology matching approaches:

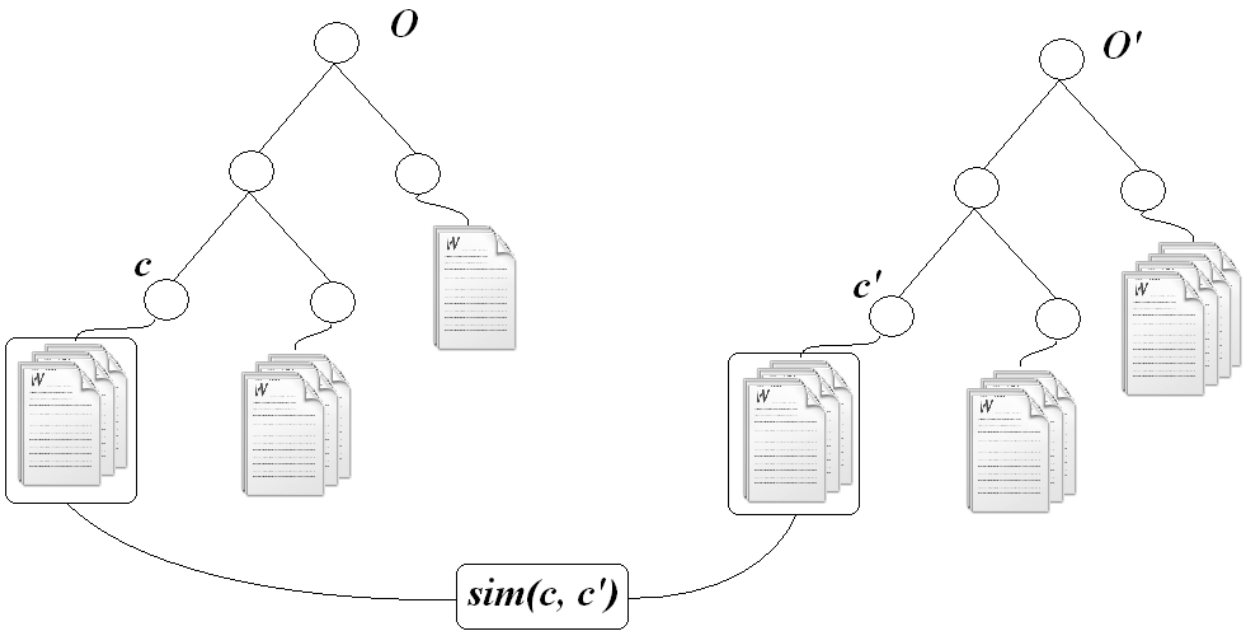
- terminological, structural, semantic, instance-based

Basic elements:

- a concept similarity measure and an algorithm which applies it, taking into account the structure.

Ontology Matching

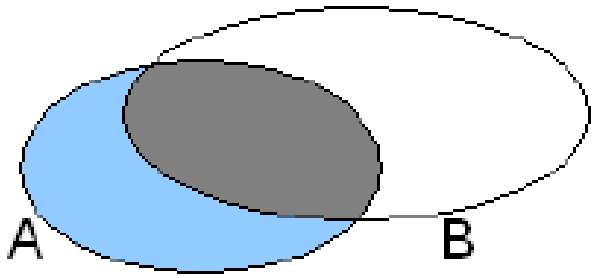
Instance-based concept similarity



The similarity of two cross-ontology concepts is assessed by the help of the instances of these concepts

-> Many possible measures.

Intersection of class instance sets

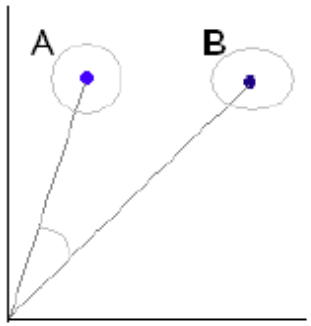


-> Same instances need to be found in both ontologies.

The cosine of the prototypes

$$sim(A,B) = s\Big(\frac{1}{|A|}\sum_{j=1}^{|A|}\mathbf{i}_j^A, \frac{1}{|B|}\sum_{k=1}^{|B|}\mathbf{i}_k^B\Big) \ ,$$

with $s(x,y)$ the cosine similarity of x and y.

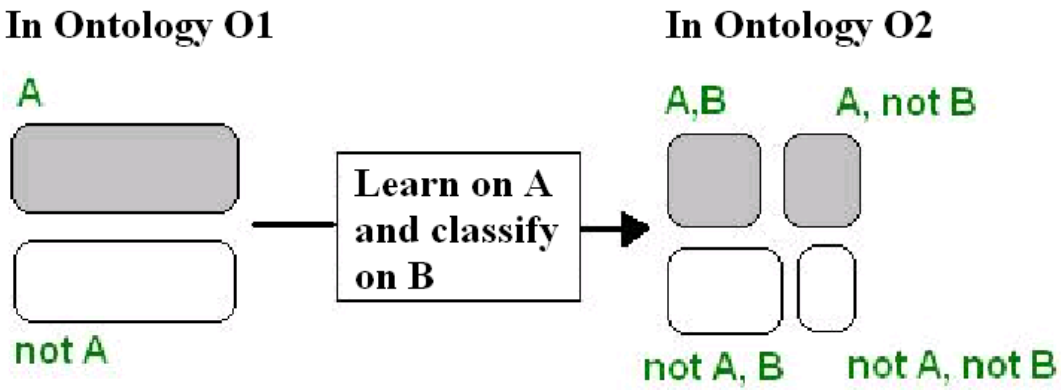


-> Flattening class structure

The Jaccard coefficient

$$Jacc(A,B) = Pr(A \cap B)/Pr(A \cup B).$$

Machine learning is used to estimate the joint probabilities.

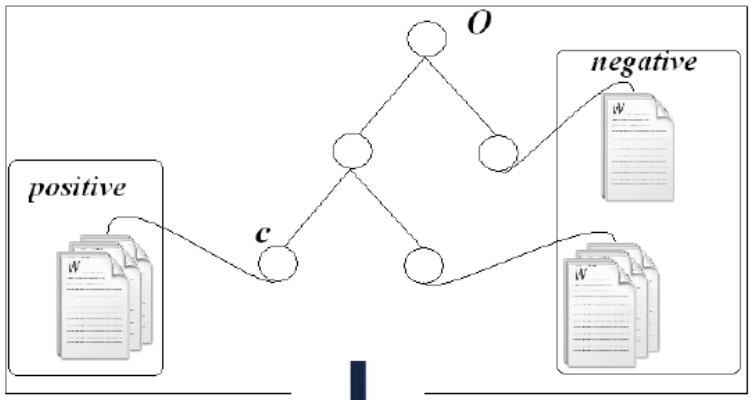


-> Insensitive to instance set intersection size

Ontology Matching

Instance-based concept similarity

Variable selection based measure

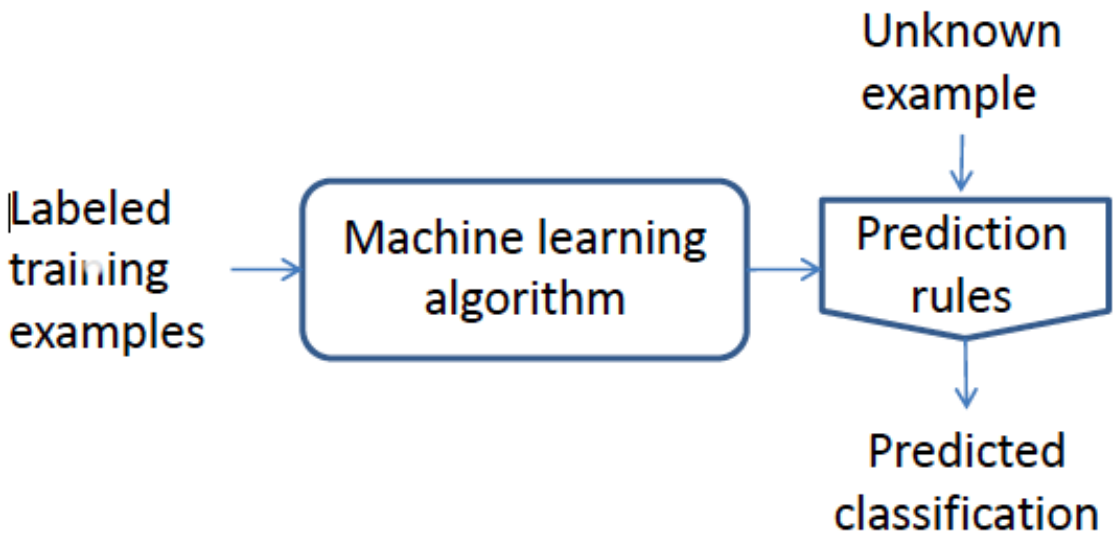


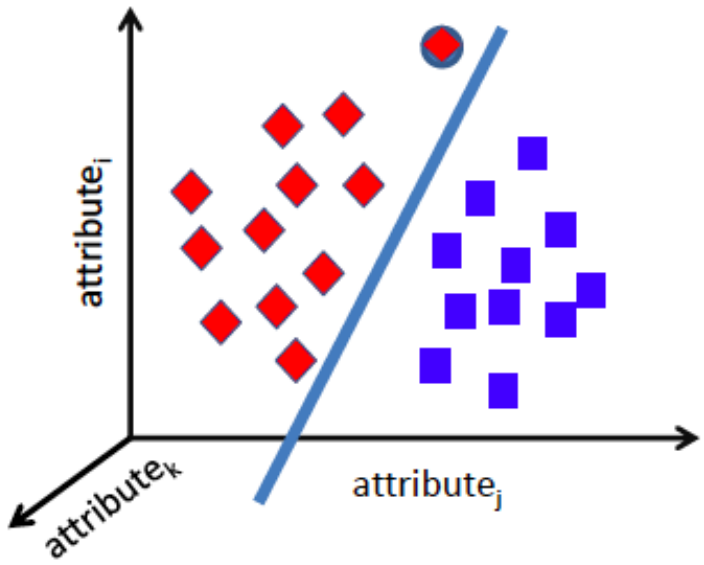
$$L(c) = (s_1^c, s_2^c, ..., s_n^c)$$

-> Time complexity is high

Ontology Matching

ML for combining similarity measures





Ontology matching	Supervised binary classification
<source entity, target entity>	Example
Similarity measures	Attribute names
Similarity values	Attribute values
Confidence value	Predicted class

Outline

Semantic Annotation

Semantic metadata are often handcrafted.

Towards an automation of this process.

ML techniques: an automatic description of a document (text, image,...) or a data entity.

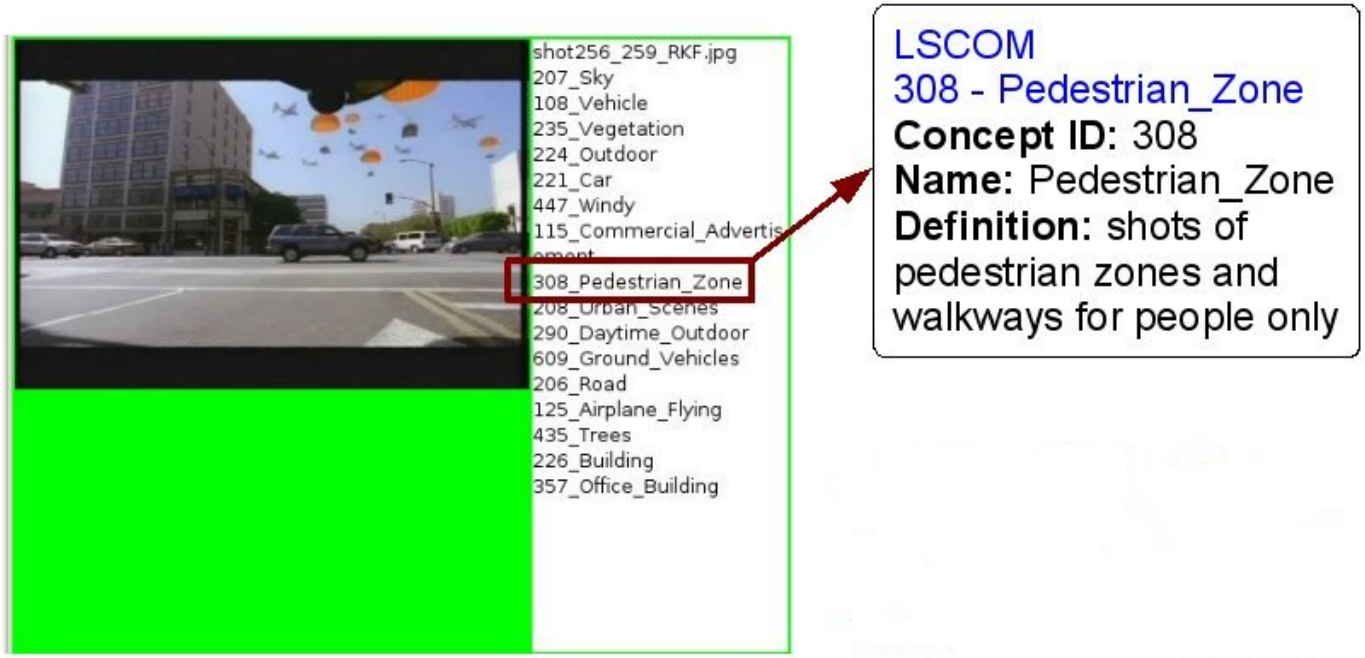
- results in a set of tags (key words that correspond to the semantic content of the document)

Semantic Annotation

=> Ontologies are learned (semi-)automatically to annotate data

<= Already constructed ontologies are used to annotate large collections of text documents, images, videos by the help of classification techniques

- Examples in multimedia: LSCOM, LabelMe



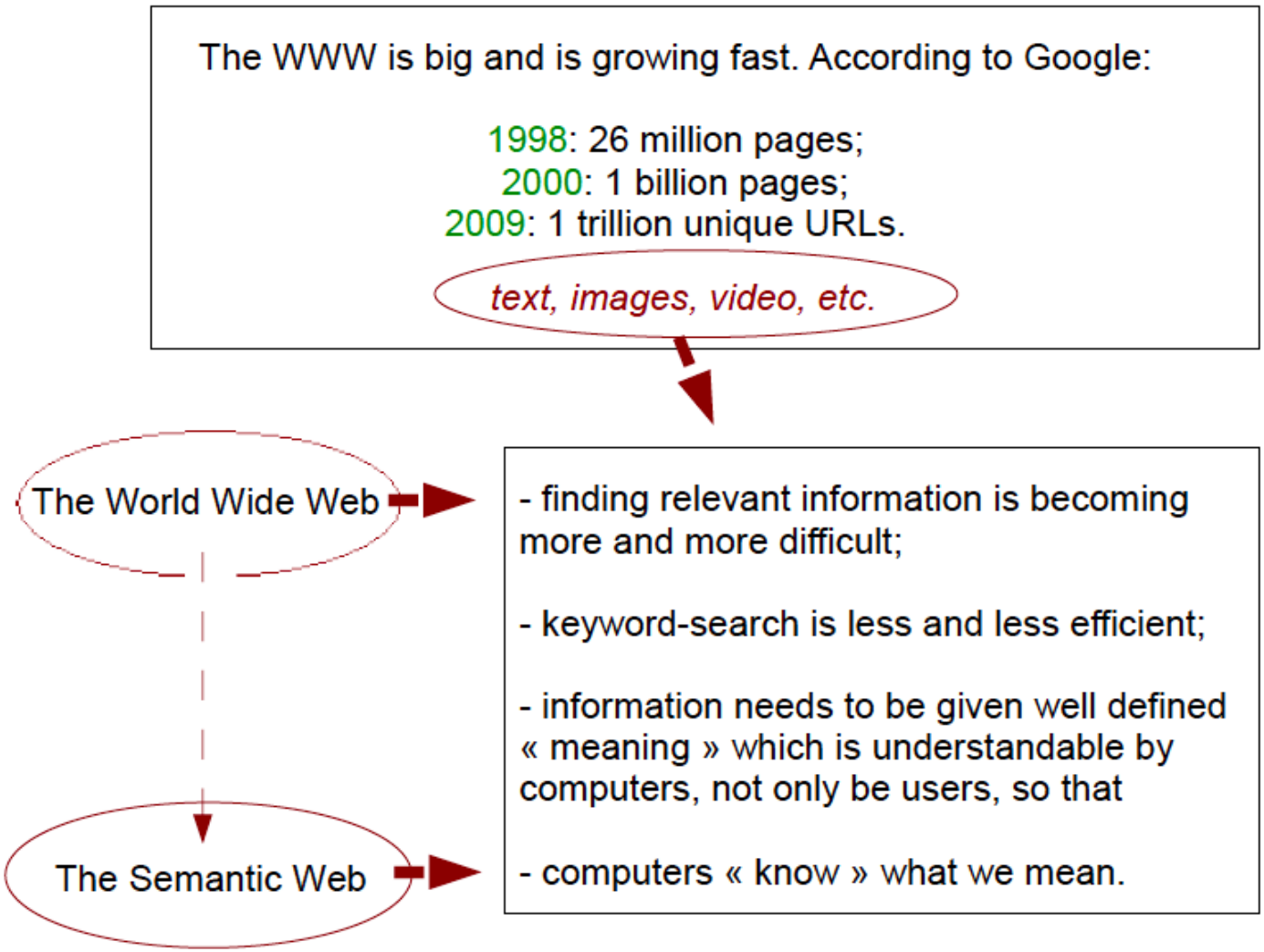
Semantic Annotation

Allow to reason over the annotations and improve them in order to describe better a document.

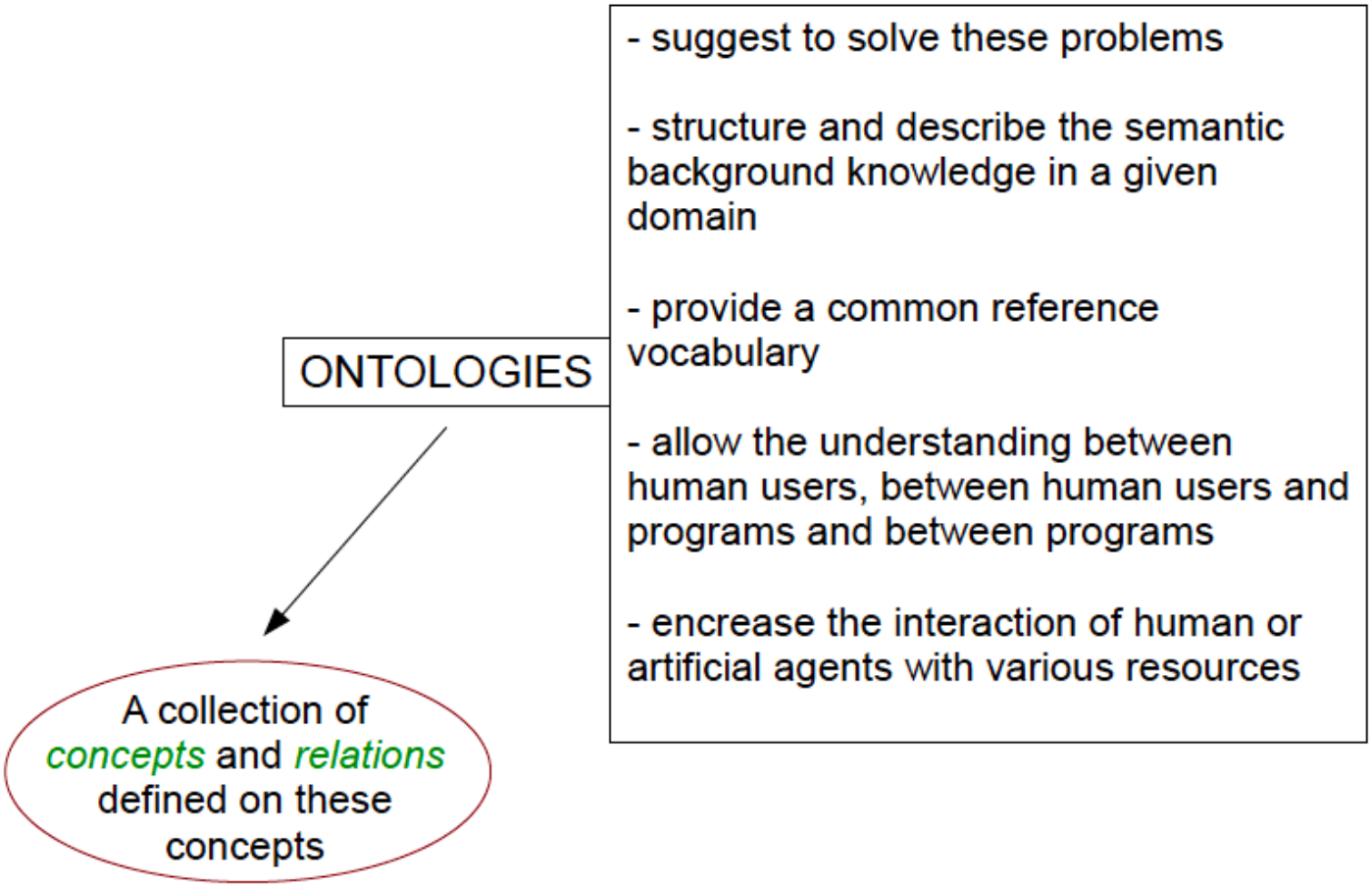
- infer new tags
- discover inconsistencies and remove incorrect tags
- use ontologies for spatial relations, context

Outline

Summary


The WWW is big and is growing fast. According to Google:
1998: 26 million pages;
2000: 1 billion pages;
2009: 1 trillion unique URLs.
text, images, video, etc.
The World Wide Web
The Semantic Web
- finding relevant information is becoming more and more difficult;
- keyword-search is less and less efficient;
- information needs to be given well defined « meaning » which is understandable by computers, not only be users, so that
- computers « know » what we mean.

Summary


ONTOLOGIES
- suggest to solve these problems
- structure and describe the semantic background knowledge in a given domain
- provide a common reference vocabulary
- allow the understanding between human users, between human users and programs and between programs
- encrease the interaction of human or artificial agents with various resources
A collection of concepts and relations defined on these concepts

Summary

Data mining is very useful to build ontologies:

extracting terms
–> constructing concepts
–> organizing concepts in hierarchies
–> defining non-hierarchical relations

Further:
–> dealing with ontology matching
–> providing automatic semantic annotation

Summary

Towards...

- An improved relation discovery (Wikipedia?)
- Ontology learning from social data
- Ontology learning across different languages

Acknowledgements and references

Parts of this course are freely inspired by the tutorial of Steffen Staab and Andreas Hotho[2], as well as by the course of Marie-Aude Aufaure[3]. The PhD thesis of Wilson Wong[4] has served as an overview of ontology learning techniques. Further sources include author's own[5].

Some further reading

W. Wong, W. Liu, and M. Bennamoun (2012): Ontology Learning from Text: A Look back and into the Future. ACM Computing Surveys, Volume 44, Issue 4, Pages 20:1-20:36.

A. Maedche and S. Staab (2001): Ontology learning for the Semantic Web. IEEE Intell. Syst. 16, 2, 72 Ð79.

A. Maedche and S. Staab (2000). The Text-to-Onto ontology learning environment. In Proceedings of the 8th International Conference on Conceptual Structures

K. Todorov, P. Geibel and K.-U. Kühnberger (2010): Mining Concept Similarities for Heterogeneous Ontologies. In: P. Perner (Ed.): Advances in Data Mining. Applications and Theoretical Aspects, ICDM 2010, pp . 86-100 . Lecture Notes in Computer Science 6171 Springer 2010.

[2] http://fr.slideshare.net/butest/semantic-web-and-machine-learning-tutorial#btnPrevious

[3] http://europa-eu-audience.typepad.com/files/sem_web_ecolia-aufaure-22mars08.pdf

[4] https://repository.uwa.edu.au

[5] e.g., http://dl.acm.org/citation.cfm?id=1880681