Bringing Together Heterogeneous Domain Ontologies via the Construction of a Common Fuzzy Knowledge Body

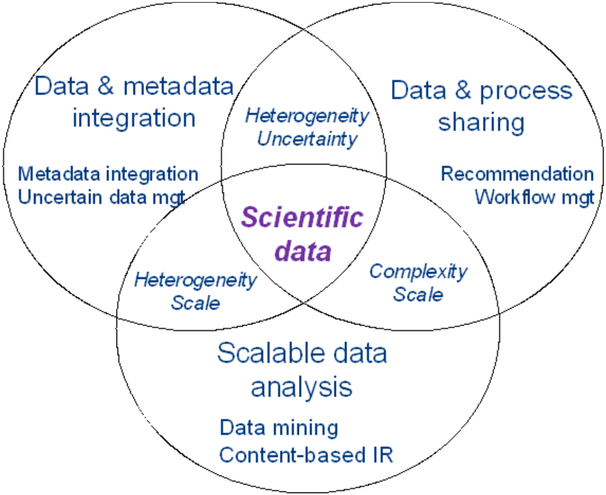
Konstantin Todorov
Peter Geibel, Céline Hudelot

todorov@lirmm.fr

ZENITH, LIRMM Montpellier
Institut für Medizinische Informatik - Charité Berlin, Germany
Ecole Centrale Paris

October 2012

ZENITH in a Nutshell



Outline

Outline

Integrating domain ontologies

Ontologies and Heterogeneity

An ontology

- describes formally and explicitly a given domain of interest o
- a semantically structured vocabulary describing data

The process of knowledge extraction and acquisition

- decentralized, highly human biased

=> Creation of ontologies that

- are (partly) complementary or (partly) redundant
- different scopes and application purposes
- mismatches in terms of syntax and terminology.

=> Ontology Heterogeneity

=> Need for knowledge and data *sharing*, *integration* and *re-contextualisation*

Bring together heterogeneous data by bridging the gap between the vocabularies that describe them.

Ontology matching addresses the heterogeneity problem

- provide a set of assertions on the relations holding between the elements of two or more heterogeneous ontologies

Building a common knowledge body: a merged ontology for a set of domain ontologies

- captures and exposes various relations holding between the concepts of the domain ontologies

The presented approach combines the use of

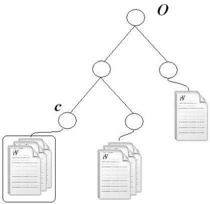
- Reference Ontology
  - Ontology matching approaches often complemented by the use of background knowledge
  - Example: weakly structured models
- Fuzzy Framework
  - Handle the inherently vague definitions of concepts; model the uncertainty of the matching results
  - Example: finding correspondence between related but different concepts

Outline

Outline

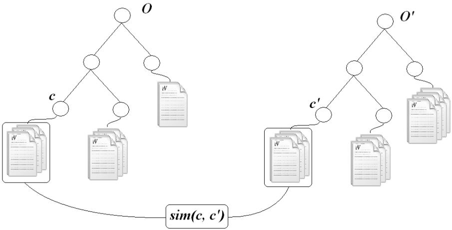
A (Populated) Ontology

- $O = \{C, is\_a, R, I, g\}$
- C is a set whose elements are called **concepts**
- is_a is a **partial order** on C
  (the existance of a hierarchical backbone is assumed)
- 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**
  (data: *text documents*, *images*, or other)
- $g : C \rightarrow 2^I$ is an injection from the set of concepts to the set of subsets of I



A matching procedure requires a (set of) **similarity measure(s)** and an **algorithm** which applies it/them.

Example: instance-based matching.



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

-> **Many possible measures**: structural, name-based, semantic, etc.

Outline

Fuzzy Sets

A fuzzy set A is defined on a given domain of objects X by the function

$$\mu_A : X \longmapsto [0,1],$$

which expresses the degree of membership of every element of X to A by assigning to each $x \in X$ a value from the interval $[0,1]$.

- The *intersection* of two fuzzy sets A and B is given by a function $T(\mu_A(x), \mu_B(x))$ referred to as a t-norm. The Gödel t-norm is defined by $T_G(a,b) = min(a,b)$.
- The *union* of two fuzzy sets A and B is given by $S(\mu_A(x), \mu_B(x))$ where S is a t-conorm. The Gödel definition is given by $S_G(a,b) = max(a,b)$.
- The *implication*, denoted $A \to B$, is given by $\mu_{A\to B}(x) = i(\mu_A(x), \mu_B(x))$ where i is an implication function. The Gödel implication is given as

$$\mu_{A\to B}(x) = \begin{cases} 1, & \text{if } \mu_A(x) \leq \mu_B(x), \\ \mu_B(x), & \text{otherwise.} \end{cases} \quad (1)$$

Outline

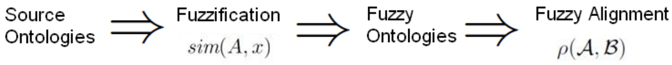
General setting

We have as an input

- a set of source ontologies $O_1, O_2, ...$
- a special, reference vocabulary $O_{ref} = (X, is\_a, R_{ref}, I_{ref}, g_{ref})$
    - The existence and suitability of a reference vocabulary: a non-trivial, highly application dependent question

The procedure:

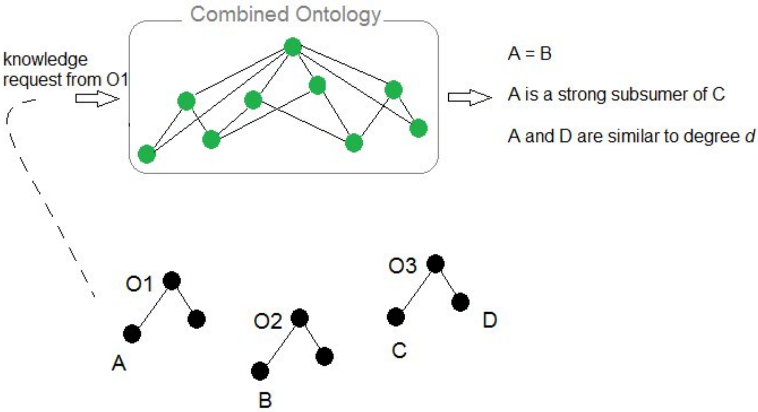
- Projection: represent each source concept as a fuzzy set on the set of reference concepts
- Union: match the source concepts to one another by the help of fuzzy set relatedness criteria



General setting

We have as an output:

- A novel *fuzzy* ontology which represents the relations holding between the concepts of the source ontologies.



Outline

Concept Fuzzification

Fuzzy representation of the source concepts

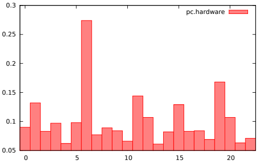
A given source concept A is represented as a fuzzy set $\mathscr{A}$ with a membership function

$$\mu_{\mathscr{A}}(x) = score_A(x), \forall x \in X, \tag{2}$$

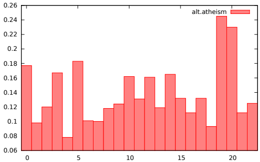
where $score_A(x)$ is the similarity score associated to a concept $x \in X$ with respect to A.

Concept Fuzzification

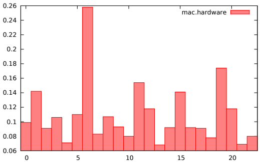
Examples



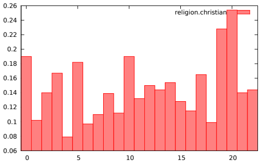
(a) Scores for *pc.hardware*



(b) Scores for *alt.atheism*



(c) Scores for *mac.hardware*



(d) Scores *religion.christian*

Figure: Fuzzy membership functions: 20NG classes scored by Wikipedia.

Algorithm: For a source ontology O_1 and a reference ontology O_{ref} :

- Leaf node of O_1: assign a score by computing crisp concept similarities to the reference concepts of O_{ref}.
- Non-leaf node of O_1: score as the maximum of the scores of its children for every $x \in X$.

Example: A has children A' and A'' and A'''. Compute:

$$score_A(x) = max\{score_{A'}(x), score_{A''}(x), score_{A'''}(x)\}, \forall x \in X. \quad (3)$$

Entailment: $\mu_{\mathscr{A}'}(x) \rightarrow \mu_{\mathscr{A}}(x) = 1$ for all x and all children $\mathscr{A}'$ of $\mathscr{A}$

Outline

Definition (Fuzzy Subsumption)

The subsumption $\mathscr{A}'$ is a $\mathscr{A}$ is defined and denoted in the following manner:

$$\textbf{is\_a}(\mathscr{A}', \mathscr{A}) = \inf_{x \in X} \mu_{\mathscr{A}' \to \mathscr{A}}(x) \tag{4}$$

Remark. Crisp subsumption is preserved: the hierarchical procedure for concept fuzzification assures that **is_a**$(A', A) = 1$ holds.

Application. Computing the fuzzy **is_a** between two concepts allows for answering a query regarding *generality* and *specificity* of cross-ontology concepts wrt a given concept.

Definition (Fuzzy Ontology)

- $\mathscr{C}$: a set of (fuzzified) concepts
- **is_a** : $\mathscr{C} \times \mathscr{C} \to [0,1]$ a fuzzy *is_a*-relationship
- $\mathscr{R}$ a set of fuzzy relations on $\mathscr{C}$ ($\mathscr{R}$ contains relations $r : \mathscr{C}^n \to [0,1]$)
- $\mathscr{X}$ a set of objects
- $\phi : \mathscr{C} \to \mathscr{F}(\mathscr{X}, [0,1])$ a function that assigns a membership function to every fuzzy concept in $\mathscr{C}$.

We require that **is_a** and ϕ are compatible, i.e., that **is_a**$(\mathscr{A}', \mathscr{A}) = \inf_x \mu_{\mathscr{A}' \to \mathscr{A}}(x)$ holds for all $\mathscr{A}', \mathscr{A} \in \mathscr{C}$. With these definitions, the quintuple

$$\mathscr{O} = (\mathscr{C}, \textbf{is\_a}, \mathscr{R}, \mathscr{X}, \phi)$$

forms a fuzzy ontology.

Building a common fuzzy ontology

Measures for Fuzzy Concept Similarity

- Basic

$$\rho_{base}(\mu_{\mathscr{A}}, \mu_{\mathscr{B}}) = \max_{x \in X} |\mu_{\mathscr{A}}(x) - \mu_{\mathscr{B}}(x)|. \quad (5)$$

- Euclidean distance

$$\rho_{diff}(\mu_{\mathscr{A}}, \mu_{\mathscr{B}}) = \|\mu_{\mathscr{A}} - \mu_{\mathscr{B}}\|_2 . \quad (6)$$

- Zadeh's partial matching index

$$\rho_{sup-min}(\mu_{\mathscr{A}}, \mu_{\mathscr{B}}) = \sup_{x \in X} T(\mu_{\mathscr{A}}(x), \mu_{\mathscr{B}}(x)). \quad (7)$$

- Jaccard coefficient

$$\rho_{jacc}(\mu_{\mathscr{A}}, \mu_{\mathscr{B}}) = \frac{\sum_x T(\mu_{\mathscr{A}}(x), \mu_{\mathscr{B}}(x))}{\sum_x S(\mu_{\mathscr{A}}(x), \mu_{\mathscr{B}}(x))}. \quad (8)$$

The union of two fuzzy concepts can be decomposed into three components:

$$S(\mathscr{A},\mathscr{B}) = (\mathscr{A}\mathscr{B}) + (\mathscr{A}-\mathscr{B}) + (\mathscr{B}-\mathscr{A}). \tag{9}$$

$$\begin{aligned}
\mathscr{A}\mathscr{B} &= T(\mathscr{A},\mathscr{B}) && // \text{ what is common to both concepts;} && (10)\\
\mathscr{A}-\mathscr{B} &= T(\mathscr{A},\neg\mathscr{B}) && // \text{ what is characteristic for A;} && (11)\\
\mathscr{B}-\mathscr{A} &= T(\mathscr{B},\neg\mathscr{A}) && // \text{ what is characteristic for B.} && (12)
\end{aligned}$$

Several merge options can be provided to the user w.r.t. the values of these three components. In case $\mathscr{A}\mathscr{B}$ is significantly larger than each of $\mathscr{A}-\mathscr{B}$ and $\mathscr{B}-\mathscr{A}$, the two concepts can be merged into their union. In case one of $\mathscr{A}-\mathscr{B}$ or $\mathscr{B}-\mathscr{A}$ is larger than the other two components, the concepts can be merged to either $\mathscr{A}$ or $\mathscr{B}$.

Outline

The Reference Ontology

Extended version of Wikipedia's main topic classifications containing more than 30 semantically related concepts.

- Population: Inex 2007 corpus
- documents directly belong to a category, or to one of its direct subcategories in the Wikipedia category tree

Among the concepts:

law (745 documents), technology (293), arts(319), society(2050), agriculture(530) social_sciences(1695) computing(1902) health(341) education(515) mathematics(1903) people(136) business(1202) science(547), history(445), politics(896), applied_sciences(1302), geography(164), chronology(303), environment(467), nature(234), humanities(537), language(427), culture(765), etc...

Outline

The Source Ontologies: LSCOM and LabelMe

- natural hazard
  - earthquake
  - natural Disasters
  - tornado
  - avalanche
  - mudslide
- conveyance
  - airplane
    - flying
    - landing
    - take-off
  - ground vehicles
    - bus
    - truck
  - boat
    - sailboat
    - boat_ship
- ...
- computers
- recreational activity
  - sports
    - baseball
    - basketball
    - football
    - soccer
    - tennis
- group
- single person
  - single person male
  - single person female
- head and shoulders

Figure: The LSCOM sub-ontology populated by TRECVID images.

The Source Ontologies: LSCOM and LabelMe

- animal
  - bird
  - fish
  - mammal
    - bear
    - monkey
    - cat
    - dog
- group
  - people
  - social
    - couple
    - crowd
    - audience
- body
  - arm
  - head
  - leg
  - torso
- ...
- atmospheric phenomenon
  - cloud
  - fog
  - snow
- conveyance
  - bus
  - vehicle
    - airplane
    - truck
    - vessel
      - boat
        - motorboat
        - gondola
- device
  - electr. device
    - computer

Figure: The WordNet/LabelMe sub-ontology.

Instances

In this case, the instances are **images** represented by **text documents** relevant to the concepts generated from:

- The names of all concepts that the image contains in its annotation
- The textual definitions of these concepts (LSCOM definitions for TRECVID images and WordNet glosses for LabelMe images)

107_Standing One or more people standing up. 227_Bus Shots of a bus. 224_Outdoor Shots of Outdoor locations. 217_Person Shots depicting a person. The face may be partially visible. 202_Crowd Shots depicting a crowd. 181_Adult Shots showing a person over the age of 18. 104_Male_Person One or more male persons. 290_Daytime_Outdoor shots that take place outdoors during the day. 316_Group We defined a group as 3-10 people. 109_Windows An opening in the wall or roof of a building or vehicle fitted with glass or other transparent material.

Using LSCOM keyword description is tricky because they depend on negation, exclusion, etc. Exemple Group: "[...]It only includes shots of 3 to 1 people, not animals, such as pets, nor animationed people, such as in previews of The Incredibles"

Results

Summary of the procedure

- Fuzzification of the LSCOM concepts.
- Fuzzification of the LabelMe concepts.
- Compute fuzzy subsumtion for every pair of concepts from the union of the fuzzified LSCOM and LabelMe concepts.
- Compute similarities between the fuzzified concepts.
- Build the fuzzy common knowledge body.

Results

Intra-ontology concepts

Concept $\mathcal{A}$: Concept $\mathcal{B}$:	LSCM:truck vs. LSCM:gr.vehicle	LSCM:sports vs. LSCM:basketball	LM:computer vs. LM:elec. device	LM:animal vs. LM:bird
is_a($\mathcal{A}, \mathcal{B}$)	1	0.007	1	0.004
is_a($\mathcal{B}, \mathcal{A}$)	0.012	1	0.011	1
is_a$^{\text{mean}}$($\mathcal{A}, \mathcal{B}$)	1	0.052	1	0.062
is_a$^{\text{mean}}$($\mathcal{B}, \mathcal{A}$)	0.326	1	0.07	1
Base Sim.	0.848	0.959	0.915	0.390
Eucl. Sim.	0.835	0.908	0.854	0.350
SupMin Sim.	0.435	0.545	0.359	0.309
Jacc. Sim.	0.870	0.814	0.733	0.399
Cosine Sim.	0.974	0.994	0.975	0.551

Figure: Examples of pairs of matched **intra-ontology** concepts (column-wise).

Results

Cross-ontology concepts

Concept $\mathcal{A}$: Concept $\mathcal{B}$:	LM:gondola vs. LSCM:boat_ship	LSCM:group vs. LM:audience	LSCM:truck vs. LM:vehicle	LSCM:truck vs. LM:conveyance
$is\_a(\mathcal{A}, \mathcal{B})$	0.016	0.006	0.022	0.022
$is\_a(\mathcal{B}, \mathcal{A})$	0.009	1	0.012	0.012
$is\_a^{mean}(\mathcal{A}, \mathcal{B})$	0.86	0.022	0.748	0.769
$is\_a^{mean}(\mathcal{B}, \mathcal{A})$	0.167	1	0.301	0.281
Base Sim.	0.72	0.78	0.58	0.58
Eucl. Sim.	0.66	0.71	0.40	0.38
SupMin Sim.	0.069	0.082	0.22	0.22
Jacc. Sim.	0.49	0.42	0.54	0.52
Cosine Sim.	0.69	0.82	0.66	0.67

Figure: Examples of pairs of matched **cross-ontology** concepts, column-wise.

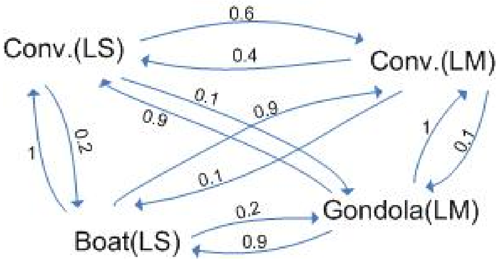
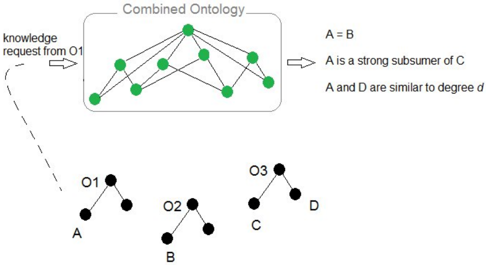


Figure: A fragment of the common fuzzy ontology of LSCOM (LS) and LabelMe (LM).

Outline

Conclusion

The proposed approach contributes to reducing metadata heterogeneity in collaborative contexts.



Future work:

-> DL framing for handling general OWL ontologies

-> large-scale integration

-> multilingualism

Thank you for listening!

Evaluation by a human driven alignment

LSCOM	**LabelMe** Depth 1	Depth 2	Depth 3
airplane takeoff	2.8: airplane	1.75: bird	1.7: conveyance
natural disasters	2.5: atm_phenom.	1.5: cloud	1.34: snow
single person	2.4: people	2.1: body	1.25: head
bus	3: bus	2.25: conveyance	2.0: vehicle
sailboat	2.9: boat	1.8: vehicle	1.8: vessel
baseball	2.2: group	2.13: social	2.0: crowd
natural hazard	2.3: atm_phenom.	1.6: snow	1.23: cloud
single person male	2.3: people	2.25: couple	2.0: head
truck	3: truck	2.0: vehicle	1.75: conveyance
ground vehicles	2.6: vehicle	2.1: truck	1.75: conveyance
football	2.5: crowd	2.3: body	2.0: social
airplane	3: airplane	2.1: vehicle	2.0: conveyance
computers	3: computer	2.14: comp_monitor	2.0: elec_device
airplane flying	2.9: airplane	2.0: atm_phenom.	1.75: conveyance
group	3: group	2.0: people	2.0: crowd
earthquake	2.25: people	1.8: crowd	1.67: fog
basketball	2.2: group	2.25: crowd	1.8: people
airplane landing	2.81: airplane	1.86: vehicle	1.75: conveyance
tennis	2.125: social	2.0: couple	2.0: body
boat	3: boat	2.2: motorboat	2.0: vessel
sports	2.63: social	2.0: people	2.0: group
single person female	2.4: body	2.0: people	2.0: couple
head and shoulder	2.6: head	2.0: torso	1.8: body
soccer	2.67: audience	2.4: crowd	2.25: leg
conveyance	3: conveyance	2.1: vehicle	1.75: vessel
avalanche	2.72: snow	2.2: atm_phenom.	1.67: body
tornado	2.8: atm_phenom.	1.89: cloud	1.3: snow
recr. activity	2.3: social	2.0: group	1.57: people
boat ship	2.9: boat	2.14: vessel	1.75: motorboat

	Recursive	Non-Recursive Depth 1	Non-Recursive Depth 2	Non-Recursive Depth 3
CRISP	0.5	0.55	0.5	**0.54**
FUZZY	**0.57**	**0.62**	**0.59**	**0.54**

Impact of the choice of a reference vocabulary

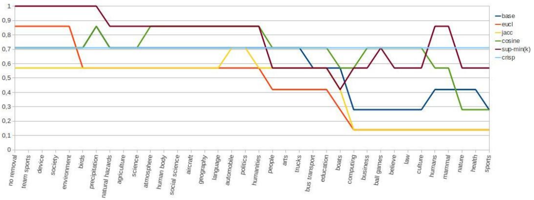
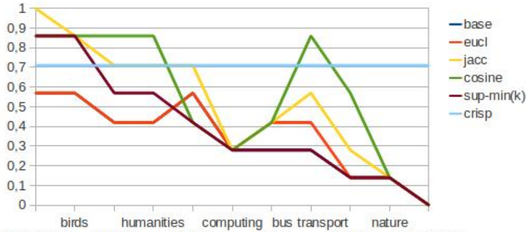


Figure: Precisions of the FuzzySim measures wrt different ref. vocabularies. On the X-axis: the different ref. vocabularies used. Below: Precisions wrt different ref. vocanbularies composed only of milestone concepts.



References

K. Todorov, P. Geibel, C. Hudelot (2011) : Building a Fuzzy Knowledge Body for Integrating Domain Ontologies. *URSW, 10th International Semantic Web Conference. ISWC 2011*.

K. Todorov, N. James, C. Hudelot (2011) : Multimedia Ontology Matching by Using Visual and Textual Modalities. *Intl Journal of Multimedia Tools and Applications. MTAP*.

K. Todorov, P. Geibel, K.-U. Kühnberger (2010) : Mining Concept Similarities for Heterogeneous Ontologies. *Advances in Data Mining*. *ICDM 2010*