Semantic Similarity Measurement for Geo-Ontologies

BENJAMIN ADAMS, NCEAS

NCEAS

UC SANTA BARBARA engineering

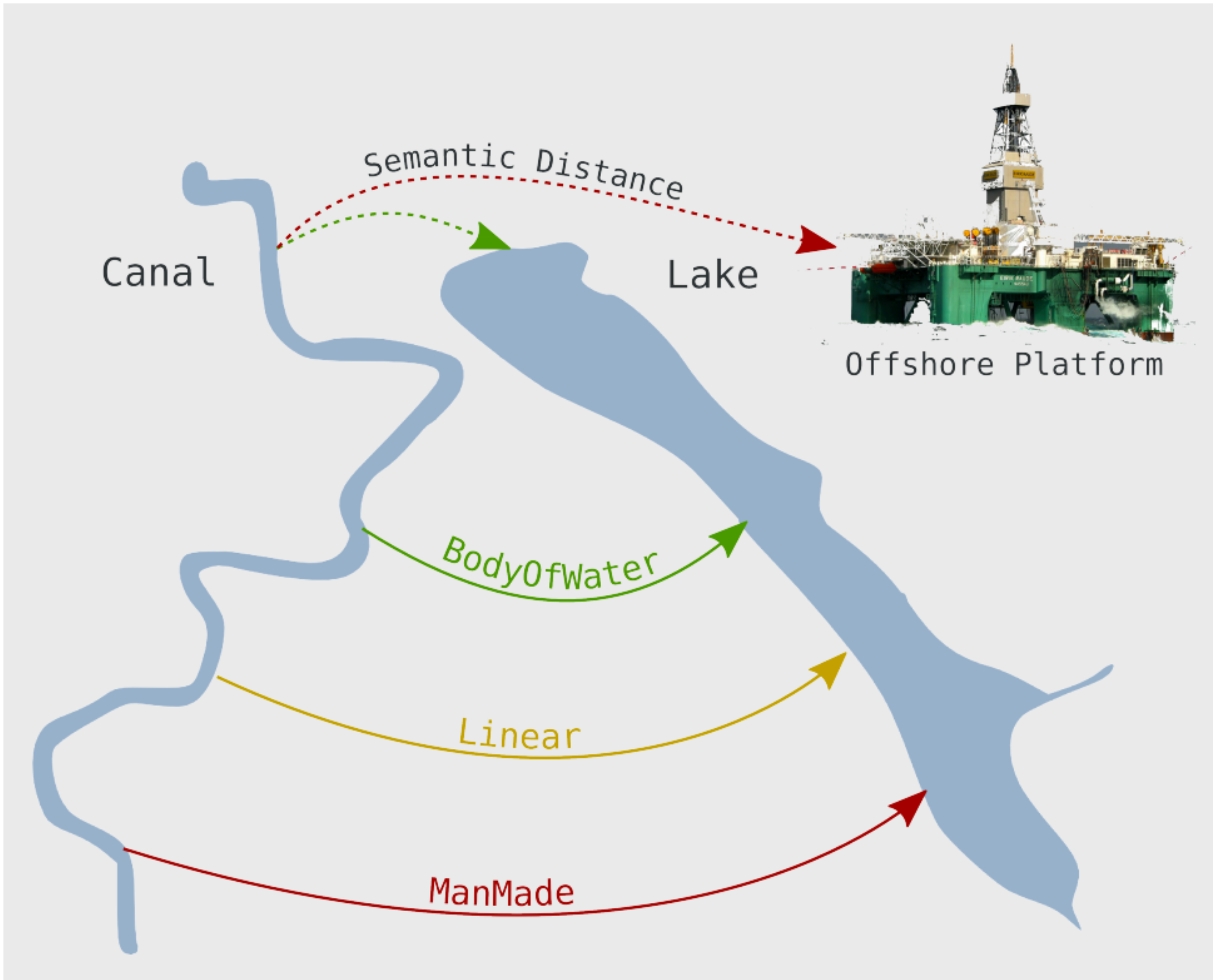
Semantic similarity

- Measuring the similarity of concepts and instances in an ontology

- Applications include:
  - Using similarity measurement to integrate information
  - Semantics based geographic information retrieval
  - Semantically enabled gazetteer services

- Focus here on concept and instance similarity in one ontology

Hydrology example



Hydrology example

$$C_s$$

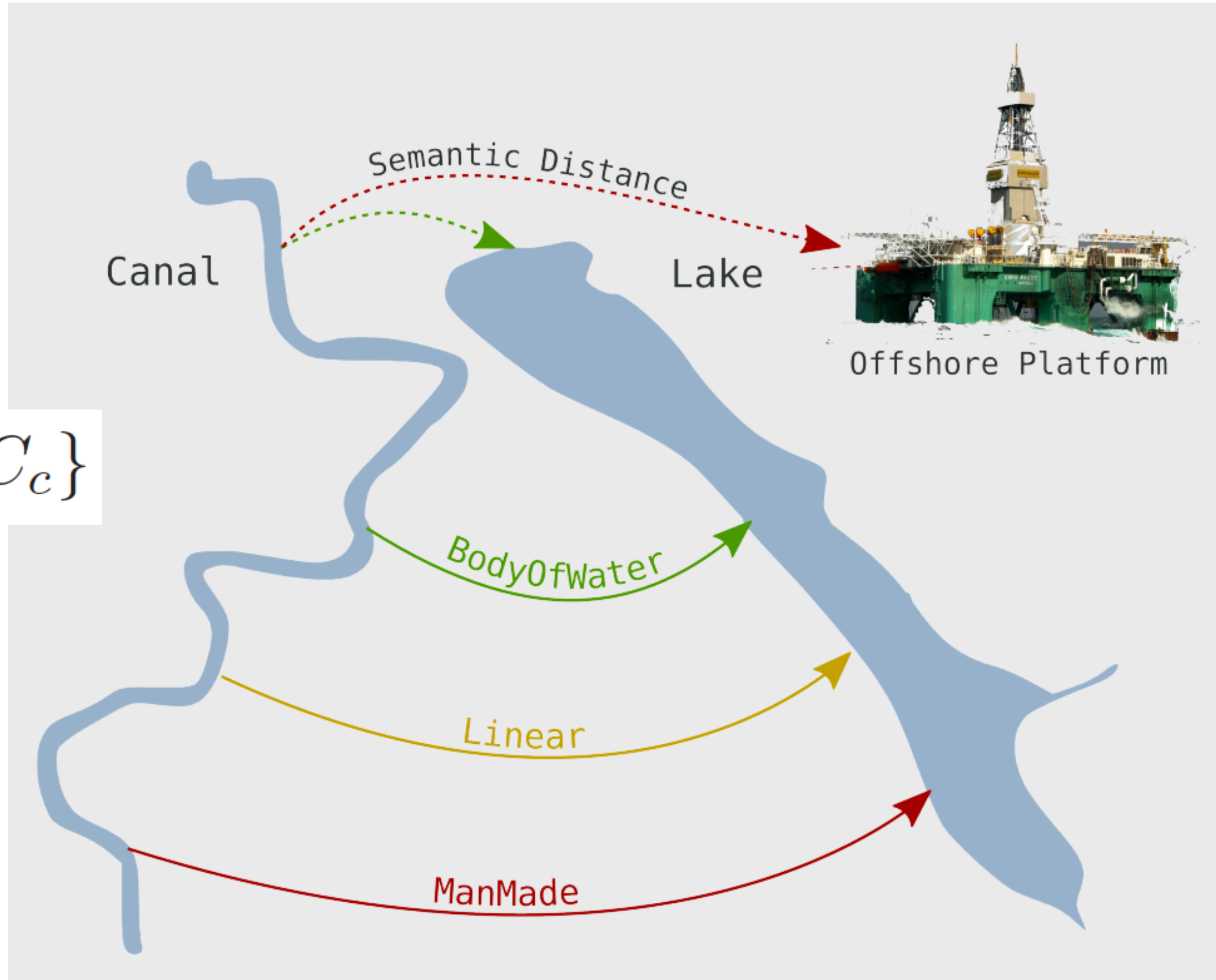
Search concept

$$C_{t_1}, \ldots, C_{t_i}$$

Target concepts

$$C_d = \{C_t \mid C_t \sqsubseteq C_c\}$$

Context of discourse



Hydrology similarity queries [from Janowicz et al. 2010]

- How similar is Canal (C_s) to River (C_t)?
- Which kind of Waterbody (C_c) is most similar to Canal (C_s)?
- What is most similar to Waterbody (C_c) ^ Artificial (C_s)?
- What is more similar to Canal (C_s), River (C_t), or Lake (C_t)?
- What are the two most similar Waterbodies (C_c) in the examined ontology?

Properties of semantic similarity

- Similarity is context-dependent [Goldstone and Son 2005]
  - A similarity measurement is meaningless without a context of discourse
- Similarity is directional and asymmetric [Tversky 1977]. CS based similarity measures tend to be symmetric
  - A lake is similar to a water body
  - A water body is similar to a lake
- Common mechanism for modeling context is to introduce *weights* on properties

Determining context weights

- How diagnostic (i.e., how relevant) is a property for similarity judgment? [Tversky 1977, Goldstone et al. 1997]
- Variability
  - If a property is shared by most entity classes being examined, it has low variability and hence less relevance
- Commonality
  - Domain of application implicitly states what properties are relevant
- Context provided by the user
  - Explicitly
  - Implicitly – e.g., inferred from a sample ranking

Several approaches to semantic similarity

- Feature overlap
- Counting transformation steps
- Finding alignments
- Computing graph-distance in a network
- Geometric spaces
- Hybrid combinations of the above
- We will just focus on a few examples.

Matching Distance Similarity Measure (MDSM)

- Extension of Tversky's ratio model [Rodriguez and Egenhofer 2004]

$$S(c_1, c_2) = \omega_p S_p(c_1, c_2) + \omega_f S_f(c_1, c_2) + \omega_a S_a(c_1, c_2)$$

$$S_t(c_1, c_2) = \frac{|C_1 \cap C_2|}{|C_1 \cap C_2| + \alpha(c_1, c_2)|C_1 C_2| + (1 - \alpha(c1, c2))|C_2 C_1|}$$

$$\alpha(c_1, c_2) = \begin{cases} \frac{d(c_1, lub)}{d(c_1, c_2)}, d(c_1, lub) \leq d(c_2, lub) \\ 1 - \frac{d(c_1, lub)}{d(c_1, c_2)}, d(c_1, lub) > d(c_2, lub) \end{cases}$$

$$d(c_1, c_2) = d(c_1, lub) + d(c_2, lub)$$

Variability: $P_t^v = 1 - \sum_{i=1}^{l} \frac{o_i}{n * l}$

Commonality: $P_t^c = \sum_{i=1}^{l} \frac{o_i}{n * l} = 1 - P_t^v$

SIM DL

- Sim DL calculates similarity of concepts and instances based on DL representation.
- Translate ontology to canonical normal form and sum of following similarities

$$sim_p(A,B) = \frac{|\{C \mid (C \sqsubseteq C_c) \wedge (C \sqsubset A) \wedge (C \sqsubset B)\}|}{|\{C \mid (C \sqsubseteq C_c) \wedge ((C \sqsubset A) \vee (C \sqsubset B))\}|}$$

Similarity (co-occurrence) of primitives

$$sim_r(R,S) = \frac{depth(lcs(R,S))}{depth(lcs(R,S)) + edge\_distance(R,S)}$$

Similarity of roles

$$sim_n(R,S) = \frac{max\_distance_n - edge\_distance(R,S)}{max\_distance_n}$$

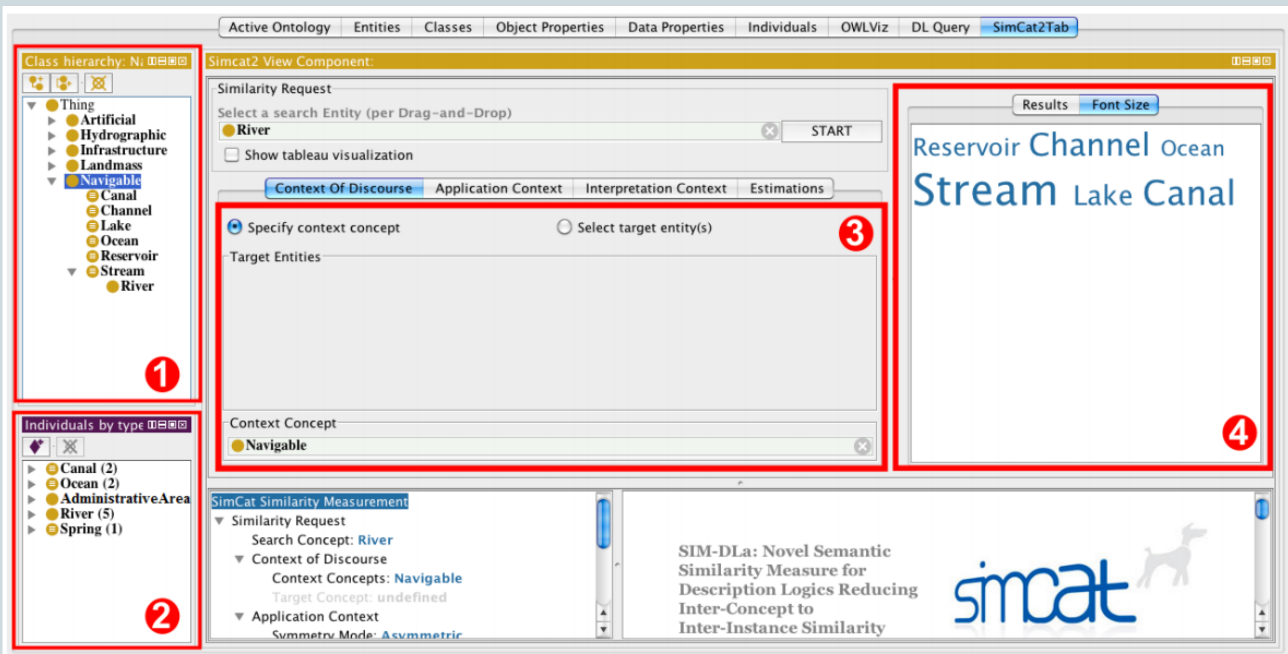
Similarity between topological or temporal relations

$$sim_{rf}(R(C),S(D)) = sim_r(R,S) * sim_o(C,D)$$

Similarity of role fillers

SIM-DL_A

- SIM-DL_A : Semantic Similarity Measurement Server
  - http://sourceforge.net/projects/sim-dl/



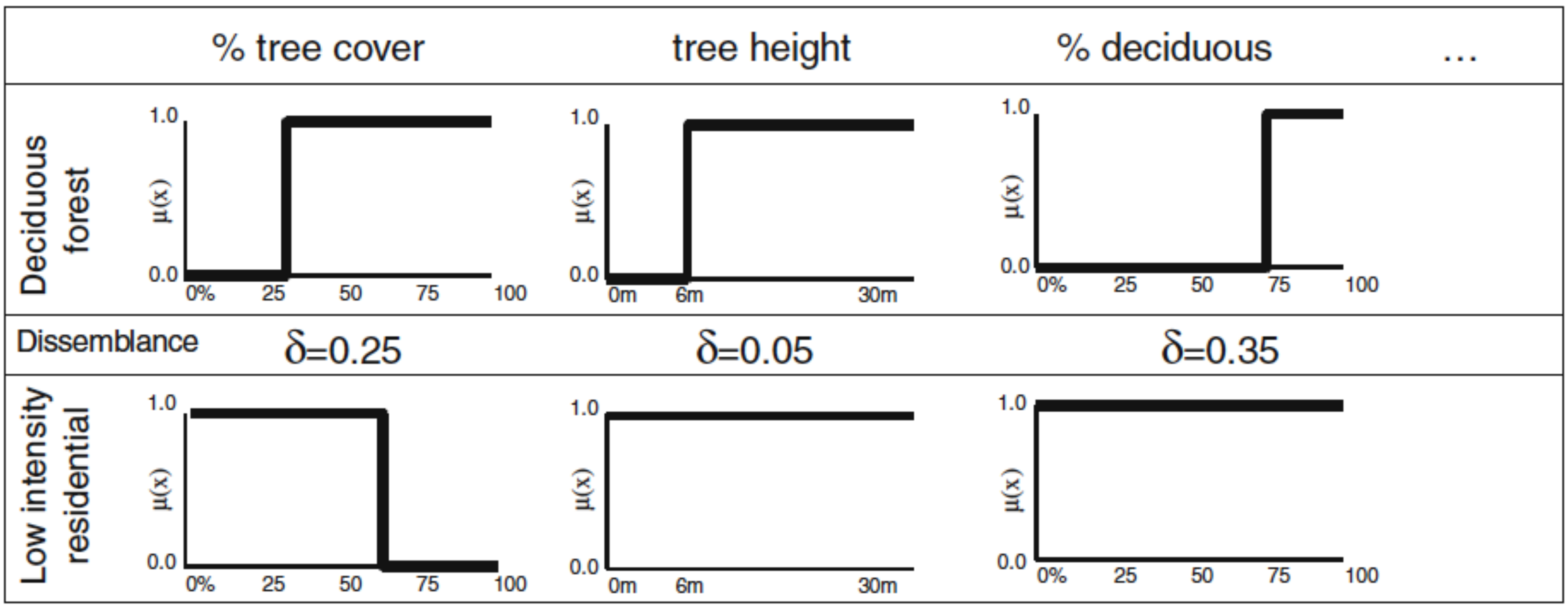
Geometric approach

- Represent semantics in a multi-dimensional space
- Semantic similarity is a function of distance in the space, e.g., Euclidean distance
- Instance similarity is distance between points
- Concept similarity
  - Distance between prototypical instances
  - Hausdorff distance between regions (facets)
  - Dissemblance index (fuzzy set interpretation)
- For DL ontologies works best with numeric datatype properties (i.e., concrete domains)

Example of geometric representation [from Ahlqvist & Shortridge 2010]

$$d(C^A C^B) = \sqrt{\sum_i^{|U|} W_i^B \delta(\mu_i^A, \mu_i^B)^2}.$$

Using a dissemblance index



References

- O. Ahlqvist and A. Shortridge (2011) Spatial and semantic dimensions of landscape heterogeneity, *Landscape Ecol*. 25:573–590.
- R. Goldstone, D. Medin, and J. Halberstadt (1997) Similarity in context. *Memory and Cognition* 25, 237–255.
- R. Goldstone and J. Son (2005) Similarity. *Cambridge Handbook of Thinking and Reasoning*, pp. 13–36.
- K. Janowicz (2006) Sim-DL: Towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval. *Proc. OTM, Part II* 1681–1692.
- K. Janowicz, B. Adams, M. Raubal (2010) Semantic Referencing - Determining Context Weights for Similarity Measurement, *GIScience 2010*, 70-84.
- K. Janowicz, M. Raubal, and W. Kuhn (2011) The Semantics of Similarity in Geographic Information Retrieval, *JOSIS* 2:29-57.
- M. A. Rodriguez and M. Egenhofer (2004) Comparing Geospatial Entity Classes: An Asymmetric and Context-dependent Similarity Measure, *IJGIS* 18(3): 229-256.
- A. Schwering (2008) Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey, *TGIS* 12(1): 5-21.
- A. Tversky (1977) Features of similarity. *Psychological Review* 84(4): 327–352.