

Semantic similarity and machine learning with ontologies

Robert Hoehndorf and Maxat Kulmanov

Semantic similarity

- We want to use *background knowledge* in ontologies to
 - ▶ determine similarity between classes,
 - ▶ instances,
 - ▶ and entities with ontology annotations

How to measure similarity?

- semantic similarity measures similarity between classes
- semantic similarity measures similarity between instances of classes
- semantic similarity measures similarity between entities *annotated* with classes
- \Rightarrow reduce all of this to similarity between classes

How to measure similarity?

What properties do we want in a similarity measure?

A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

How to measure similarity?

What properties do we want in a similarity measure?

A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

- non-negative: $sim(x, y) \geq 0$ for all x, y

How to measure similarity?

What properties do we want in a similarity measure?

A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

- non-negative: $sim(x, y) \geq 0$ for all x, y
- symmetric: $sim(x, y) = sim(y, x)$

How to measure similarity?

What properties do we want in a similarity measure?

A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

- non-negative: $sim(x, y) \geq 0$ for all x, y
- symmetric: $sim(x, y) = sim(y, x)$
- reflexive: $sim(x, x) = \max_D$

How to measure similarity?

What properties do we want in a similarity measure?

A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

- non-negative: $sim(x, y) \geq 0$ for all x, y
- symmetric: $sim(x, y) = sim(y, x)$
- reflexive: $sim(x, x) = \max_D$
 - ▶ weaker form: $sim(x, x) > sim(x, y)$ for all $x \neq y$

How to measure similarity?

What properties do we want in a similarity measure?

A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

- non-negative: $sim(x, y) \geq 0$ for all x, y
- symmetric: $sim(x, y) = sim(y, x)$
- reflexive: $sim(x, x) = \max_D$
 - ▶ weaker form: $sim(x, x) > sim(x, y)$ for all $x \neq y$
- $sim(x, x) > sim(x, y)$ for $x \neq y$

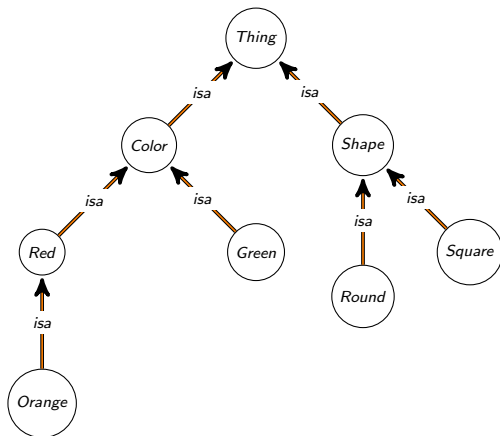
How to measure similarity?

What properties do we want in a similarity measure?

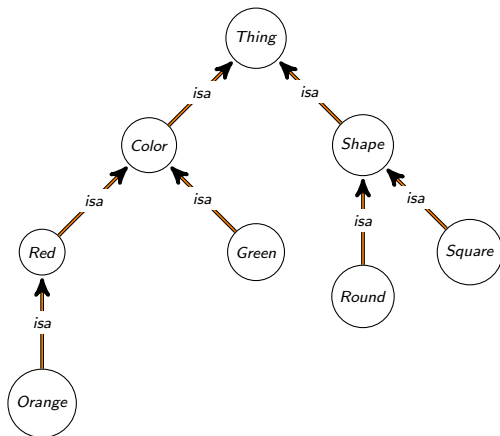
A function $sim : D \times D$ is a similarity on D if, for all $x, y \in D$, the function sim is:

- non-negative: $sim(x, y) \geq 0$ for all x, y
- symmetric: $sim(x, y) = sim(y, x)$
- reflexive: $sim(x, x) = \max_D$
 - ▶ weaker form: $sim(x, x) > sim(x, y)$ for all $x \neq y$
- $sim(x, x) > sim(x, y)$ for $x \neq y$
- sim is a *normalized* similarity measure if it has values in $[0, 1]$

How to measure similarity?

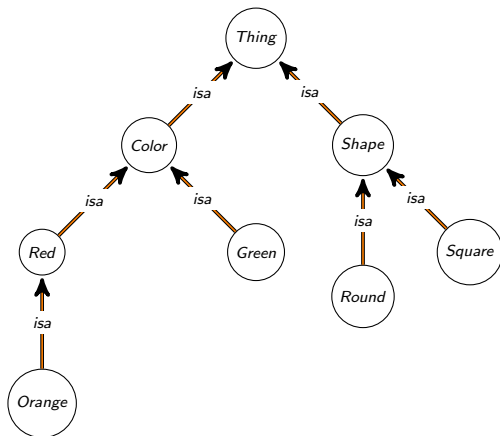


How to measure similarity?



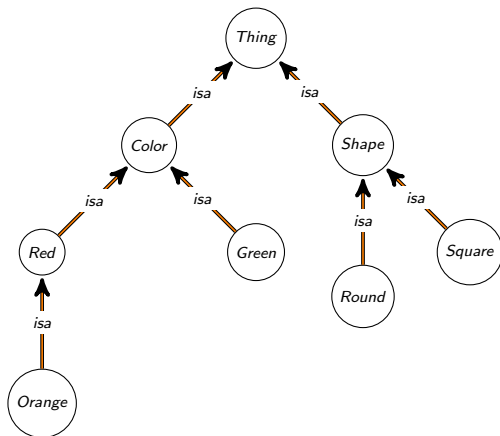
- distance on shortest path (Rada *et al.*, 1989)

How to measure similarity?



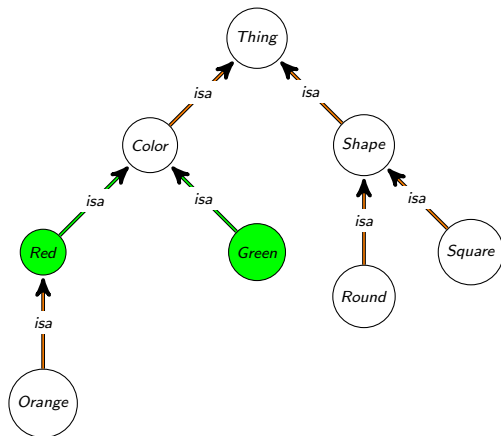
- distance on shortest path (Rada *et al.*, 1989)
- $dist_{Rada}(u, v) = sp(u, isa, v)$

How to measure similarity?



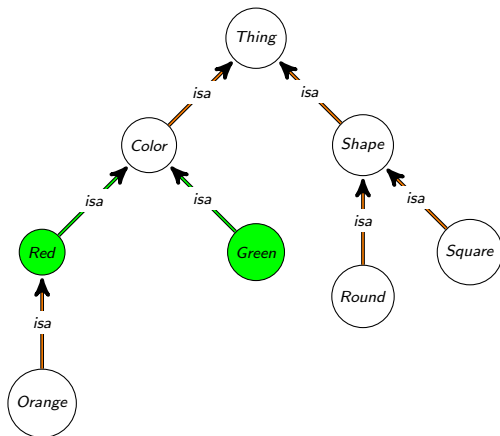
- distance on shortest path (Rada *et al.*, 1989)
- $dist_{Rada}(u, v) = sp(u, isa, v)$
- $sim_{Rada}(u, v) = \frac{1}{dist_{Rada}(u, v) + 1}$

How to measure similarity?



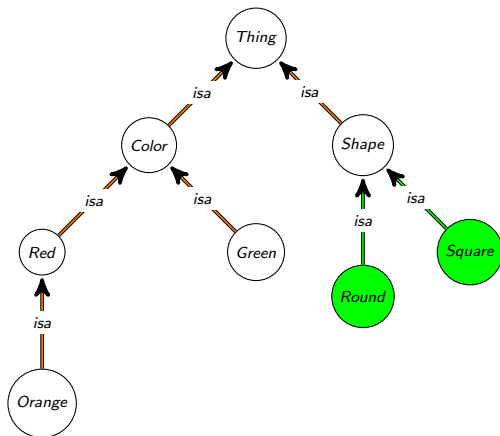
- distance on shortest path

How to measure similarity?



- distance on shortest path
- $\text{distance}(\text{green}, \text{red}) = 2$
- $\text{sim}_{\text{Rada}}(\text{green}, \text{red}) = \frac{1}{3}$

How to measure similarity?



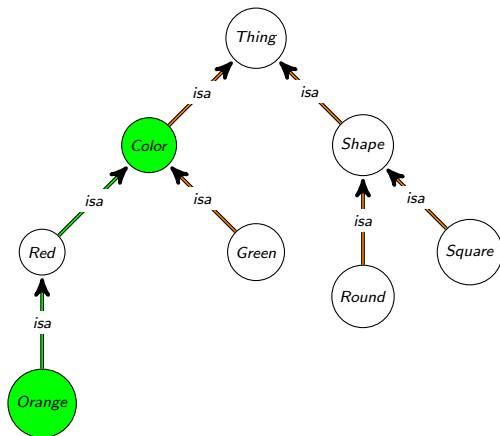
- distance on shortest path

- $\text{distance}(\text{square}, \text{round}) = 2$

-

$$\text{sim}_{\text{Rada}}(\text{square}, \text{round}) = \frac{1}{3}$$

How to measure similarity?



- distance on shortest path

- $\text{distance}(\text{orange}, \text{color}) = 2$

-

$$\text{sim}_{\text{Rada}}(\text{orange}, \text{color}) = \frac{1}{3}$$

How to measure similarity?

- shortest path is not always intuitive

How to measure similarity?

- shortest path is not always intuitive
- we need a way to determine *specificity* of a class
 - ▶ number of ancestors
 - ▶ number of children
 - ▶ information content

How to measure similarity?

- shortest path is not always intuitive
- we need a way to determine *specificity* of a class
 - ▶ number of ancestors
 - ▶ number of children
 - ▶ information content
- *density* of a branch in the ontology
 - ▶ number of siblings
 - ▶ information content

How to measure similarity?

- shortest path is not always intuitive
- we need a way to determine *specificity* of a class
 - ▶ number of ancestors
 - ▶ number of children
 - ▶ information content
- *density* of a branch in the ontology
 - ▶ number of siblings
 - ▶ information content
- account for different edge types
 - ▶ non-uniform edge weighting

How to measure similarity

- term specificity measure $\sigma : \mathcal{C} \mapsto \mathbb{R}$:
 - ▶ $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$

How to measure similarity

- term specificity measure $\sigma : C \mapsto \mathbb{R}$:

- ▶ $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$

- intrinsic:

- ▶ $\sigma(x) = f(\text{depth}(x))$

- ▶ $\sigma(x) = f(A(x))$ (for ancestors $A(x)$)

- ▶ $\sigma(x) = f(D(x))$ (for descendants $D(x)$)

- ▶ many more, e.g., Zhou et al.:

$$\sigma(x) = k \cdot \left(1 - \frac{\log |D(x)|}{\log |C|}\right) + (1 - k) \frac{\log \text{depth}(x)}{\log \text{depth}(G_T)}$$

How to measure similarity

- term specificity measure $\sigma : C \mapsto \mathbb{R}$:

- ▶ $x \sqsubseteq y \rightarrow \sigma(x) \geq \sigma(y)$

- intrinsic:

- ▶ $\sigma(x) = f(\text{depth}(x))$

- ▶ $\sigma(x) = f(A(x))$ (for ancestors $A(x)$)

- ▶ $\sigma(x) = f(D(x))$ (for descendants $D(x)$)

- ▶ many more, e.g., Zhou et al.:

$$\sigma(x) = k \cdot \left(1 - \frac{\log |D(x)|}{\log |C|}\right) + (1 - k) \frac{\log \text{depth}(x)}{\log \text{depth}(G_T)}$$

- extrinsic:

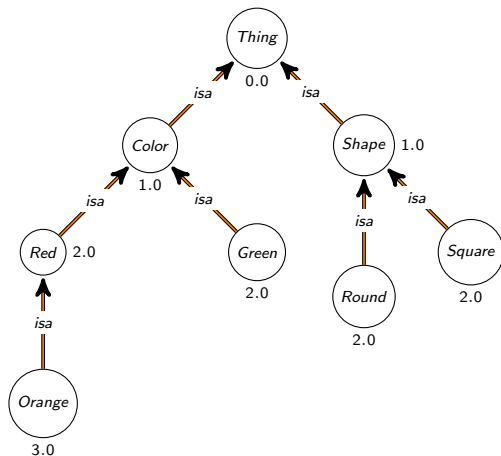
- ▶ $\sigma(x)$ defined as a function of instances (or annotations) I

- ▶ note: the number of instances monotonically decreases with increasing depth in taxonomies

- ▶ Resnik 1995: $e/C_{\text{Resnik}}(x) = -\log p(x)$ (with $p(x) = \frac{|I(x)|}{|I|}$)

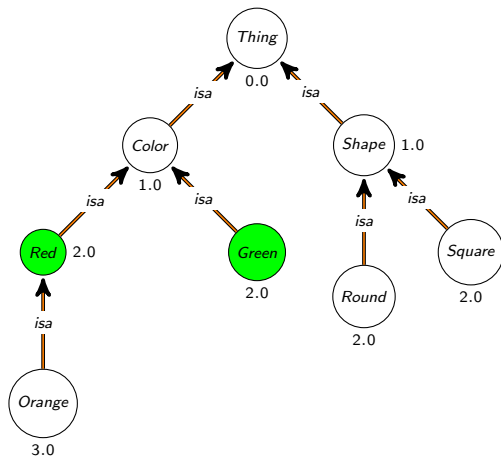
- ▶ in biology, one of the most popular specificity measure when annotations are present

How to measure similarity?



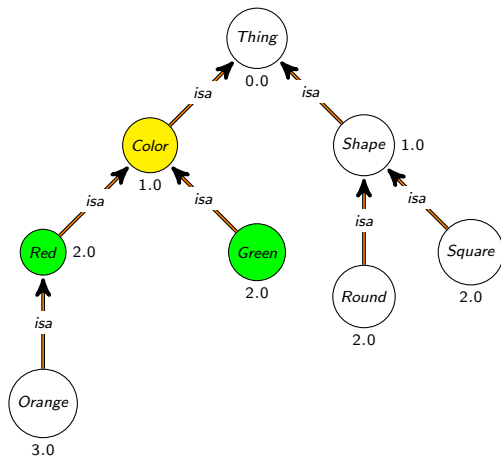
- Resnik 1995:
similarity between x
and y is the
information content
of the *most
informative common
ancestor*

How to measure similarity?



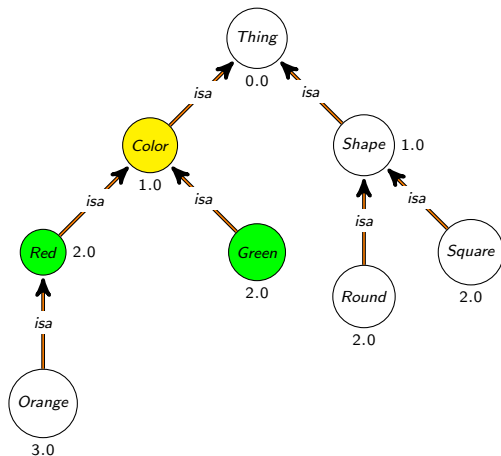
- Resnik 1995:
similarity between x
and y is the
information content
of the *most
informative common
ancestor*

How to measure similarity?



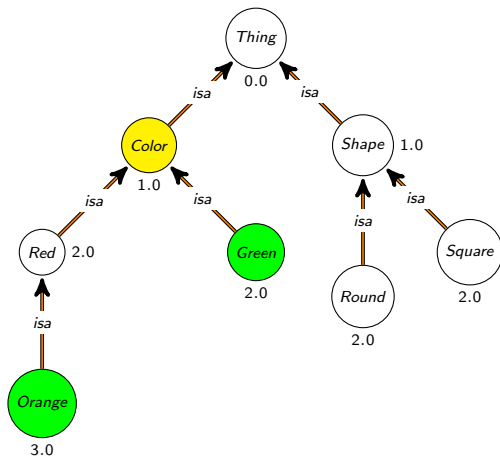
- Resnik 1995:
similarity between x
and y is the
information content
of the *most
informative common
ancestor*

How to measure similarity?



- Resnik 1995:
similarity between x and y is the information content of the *most informative common ancestor*
- $$\text{sim}_{\text{Resnik}}(\text{Green}, \text{Red}) = 1.0$$

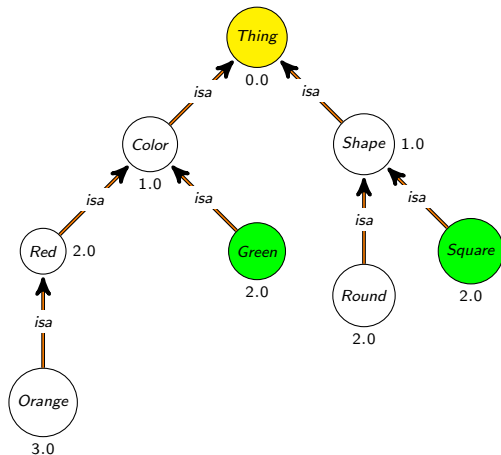
How to measure similarity?



- Resnik 1995:
similarity between x and y is the information content of the *most informative common ancestor*

- $$\text{sim}_{\text{Resnik}}(\text{Green}, \text{Orange}) = 1.0$$

How to measure similarity?

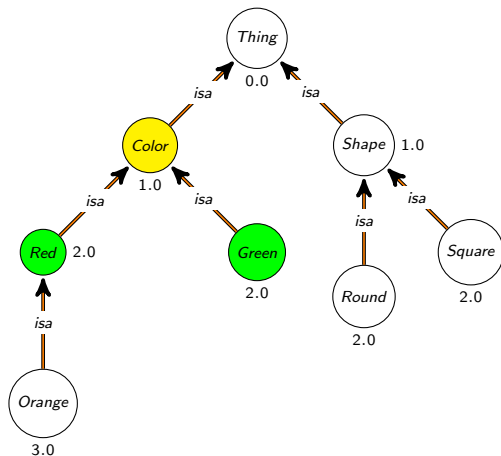


- Resnik 1995:
similarity between x and y is the information content of the *most informative common ancestor*
- $sim_{Resnik}(Square, Orange)$
0.0

How to measure similarity?

- (Red, Green) and (Orange, Green) have the same similarity
- need to incorporate the specificity of the compared classes

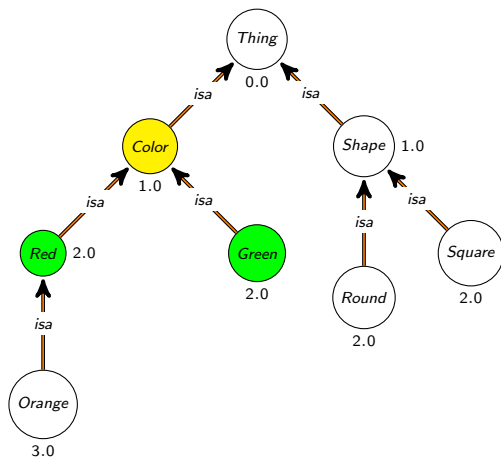
How to measure similarity?



- Lin 1998:

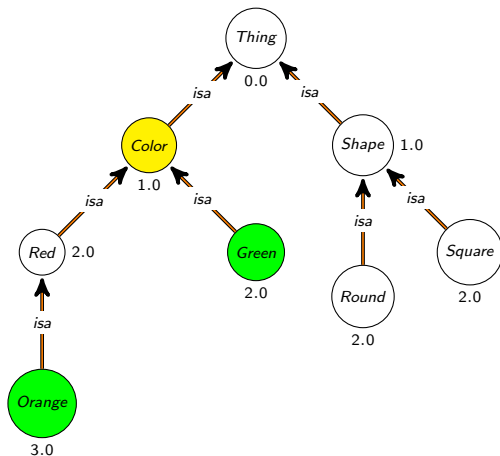
$$sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x, y))}{IC(x) + IC(y)}$$

How to measure similarity?



- Lin 1998:
$$sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x, y))}{IC(x) + IC(y)}$$
- $sim_{Lin}(Green, Red) = 0.5$

How to measure similarity?



- Lin 1998:

$$sim_{Lin}(x, y) = \frac{2 \cdot IC(MICA(x, y))}{IC(x) + IC(y)}$$

- $sim_{Lin}(Green, Orange) = 0.4$

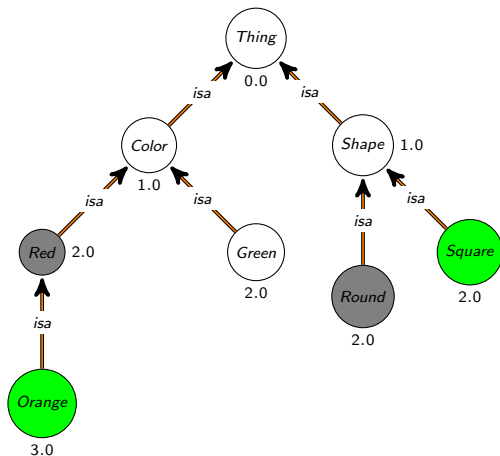
How to measure similarity?

- many(!) others:
 - ▶ Jiang & Conrath 1997
 - ▶ Mazandu & Mulder 2013
 - ▶ Schlicker et al. 2009
 - ▶ ...

How to measure similarity?

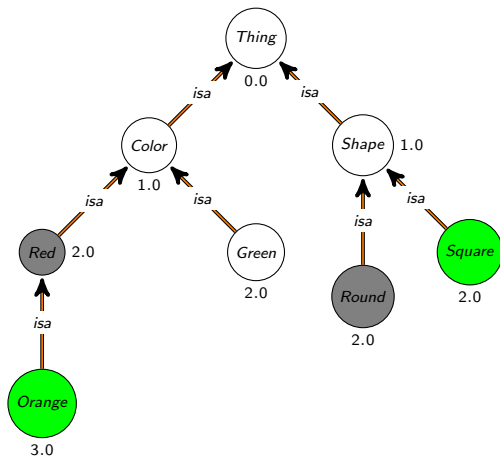
- we only looked at comparing pairs of classes
- mostly, we want to compare *sets* of classes
 - ▶ set of GO annotations
 - ▶ set of signs and symptoms
 - ▶ set of phenotypes
- two approaches:
 - ▶ compare each class individually, then merge
 - ▶ directly set-based similarity measures

How to measure similarity?



- similarity between a square-and-orange thing and a round-and-red thing

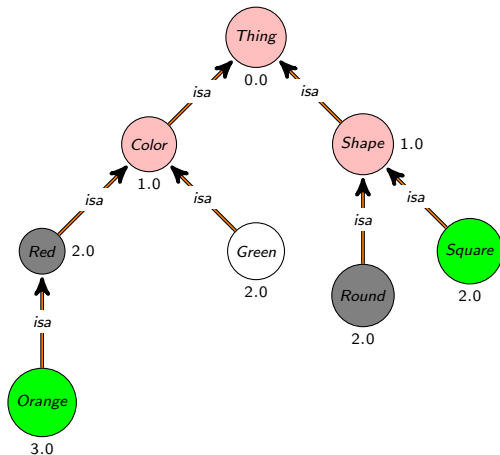
How to measure similarity?



- similarity between a square-and-orange thing and a round-and-red thing
- Pesquita et al., 2007:

$$\text{simGIC}(X, Y) = \frac{\sum_{c \in A(X) \cap A(Y)} IC(c)}{\sum_{c \in A(X) \cup A(Y)} IC(c)}$$

How to measure similarity?



- similarity between a square-and-orange thing and a round-and-red thing
- Pesquita et al., 2007:
$$simGIC(X, Y) = \frac{\sum_{c \in A(X) \cap A(Y)} IC(c)}{\sum_{c \in A(X) \cup A(Y)} IC(c)}$$
- $simGIC(so, rr) = \frac{2}{11}$

How to measure similarity?

- alternatively: use different merging strategies
- common: average, maximum, **best-matching average**
 - ▶ Average: $sim_A(X, Y) = \frac{\sum_{x \in X} \sum_{y \in Y} sim(x, y)}{|X| \times |Y|}$
 - ▶ Max average: $sim_{MA}(X, Y) = \frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} sim(x, y)$
 - ▶ Best match average: $sim_{BMA}(X, Y) = \frac{sim_{MA}(X, Y) + sim_{MA}(Y, X)}{2}$

How to measure similarity?

- Semantic Measures Library:
 - ▶ comprehensive Java library
 - ▶ <http://www.semantic-measures-library.org/>
- R packages: GOSim, GOSemSim, HPOSim, LSAfun, ontologySimilarity,...
- Python: sematch, fastsemsim (GO only)

Applications of semantic similarity

- no obvious choice of similarity measure
- depends on application
 - ▶ e.g., predicting PPIs in different organisms through similarity may benefit from a different similarity measure!
- different similarity measures may react differently to biases in data
- needs some testing and experience

Applications of semantic similarity

Recommendations:

- use Resnik's information content measure
- use Resnik's similarity
- use Best Match Average
- use the full ontology
- classify your ontology using an automated reasoner before applying semantic similarity
 - ▶ although many ontologies come pre-classified
- \Rightarrow but there are many exceptions
 - ▶ similar location \Rightarrow use location subset of GO
 - ▶ developmental phenotypes \Rightarrow use developmental branch of phenotype ontology