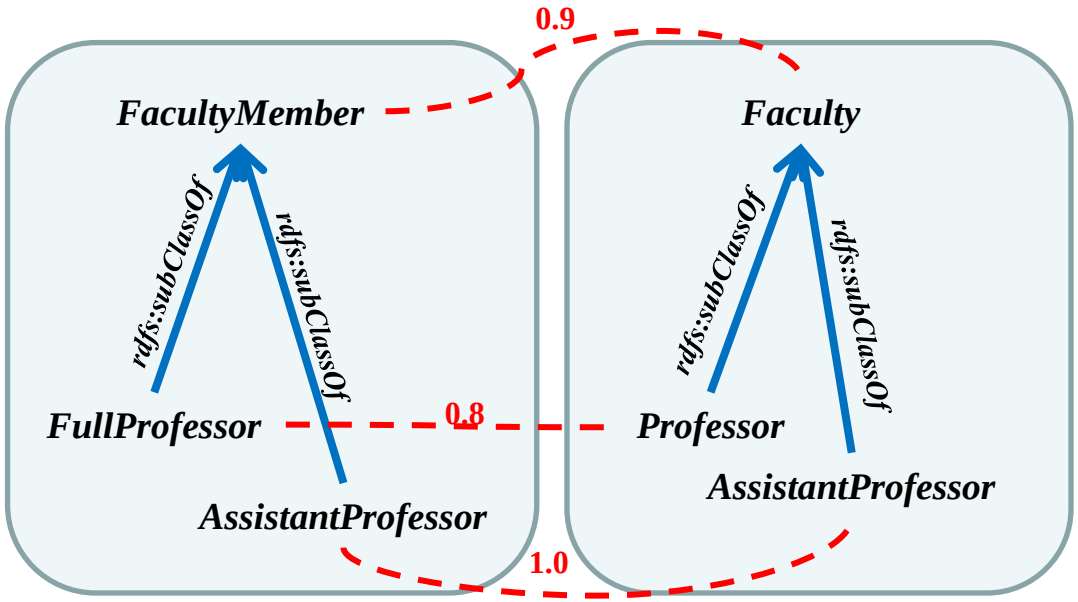
An Empirical Study of Vocabulary Relatedness and Its Application to Recommender Systems

Gong Cheng, Saisai Gong, Yuzhong Qu
State Key Laboratory for Novel Software Technology, Nanjing University, China
gcheng@nju.edu.cn

Presented at ISWC2011

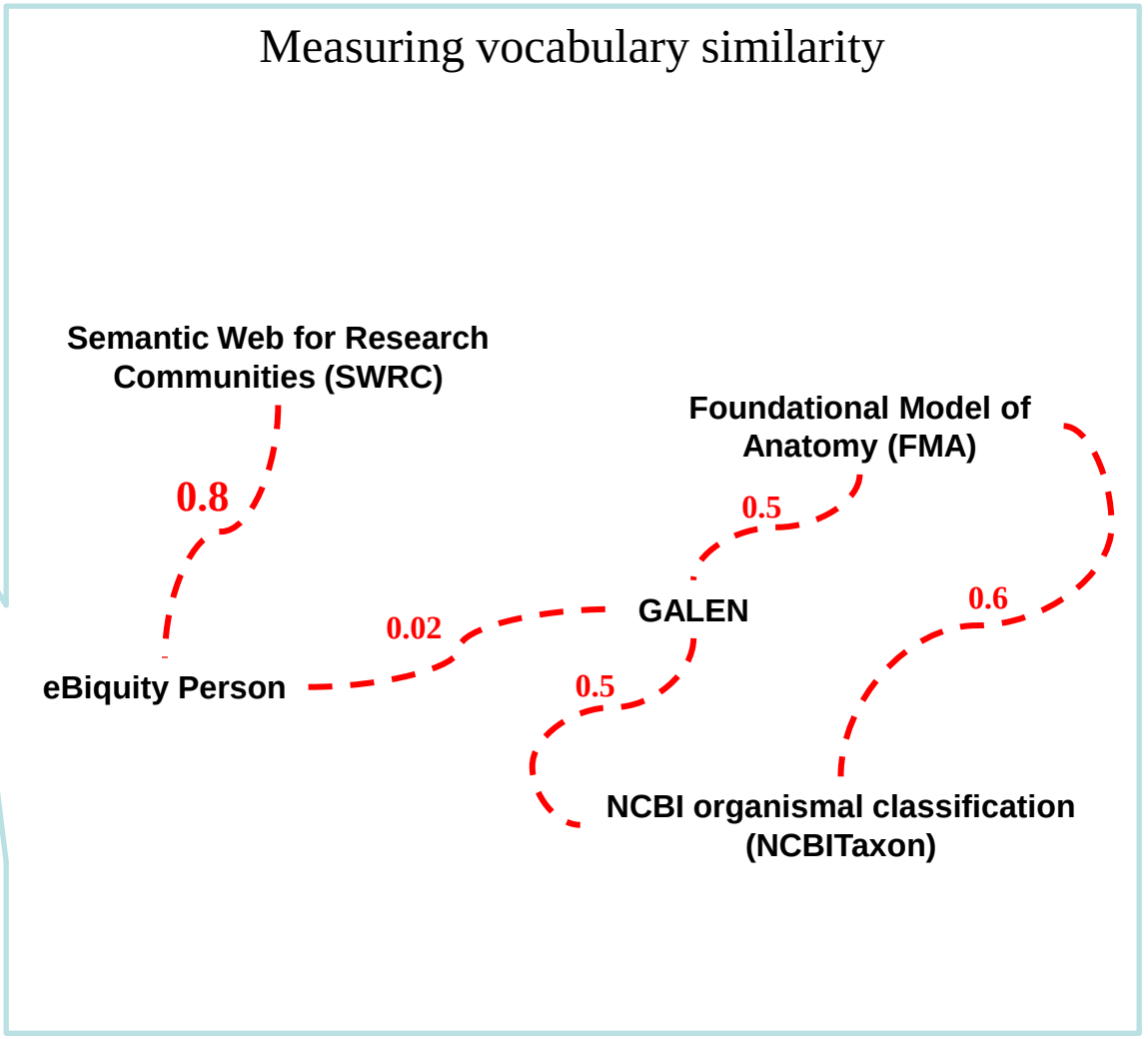
Vocabulary matching

Measuring term similarity



Vocabulary distance

Vocabulary matching

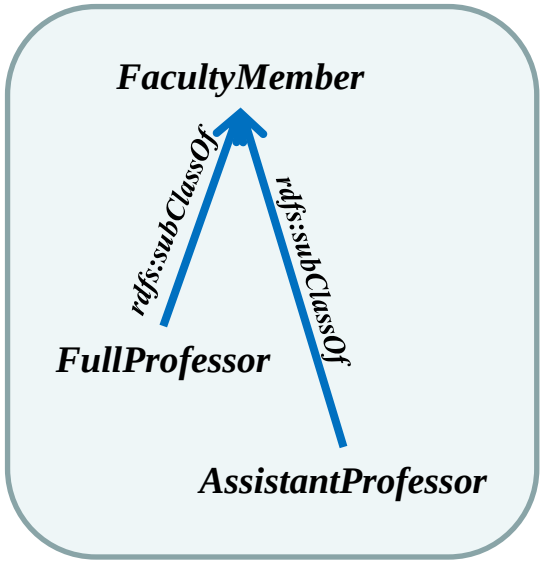


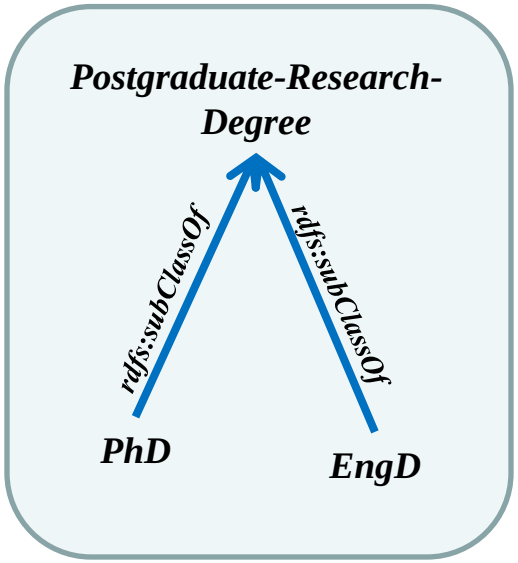
Vocabulary relatedness

Vocabulary distance

Vocabulary matching

Measuring vocabulary relatedness





not that similar, but somewhat related

Contributions

- How to measure vocabulary relatedness?
  - 6 measures, from 4 aspects
- How about vocabulary relatedness in real-life cases?
  - Empirical analysis of 2,996 vocabularies and other 4 billion RDF triples
- Where to apply vocabulary relatedness?
  - Post-selection vocabulary recommendation in vocabulary search

Outline

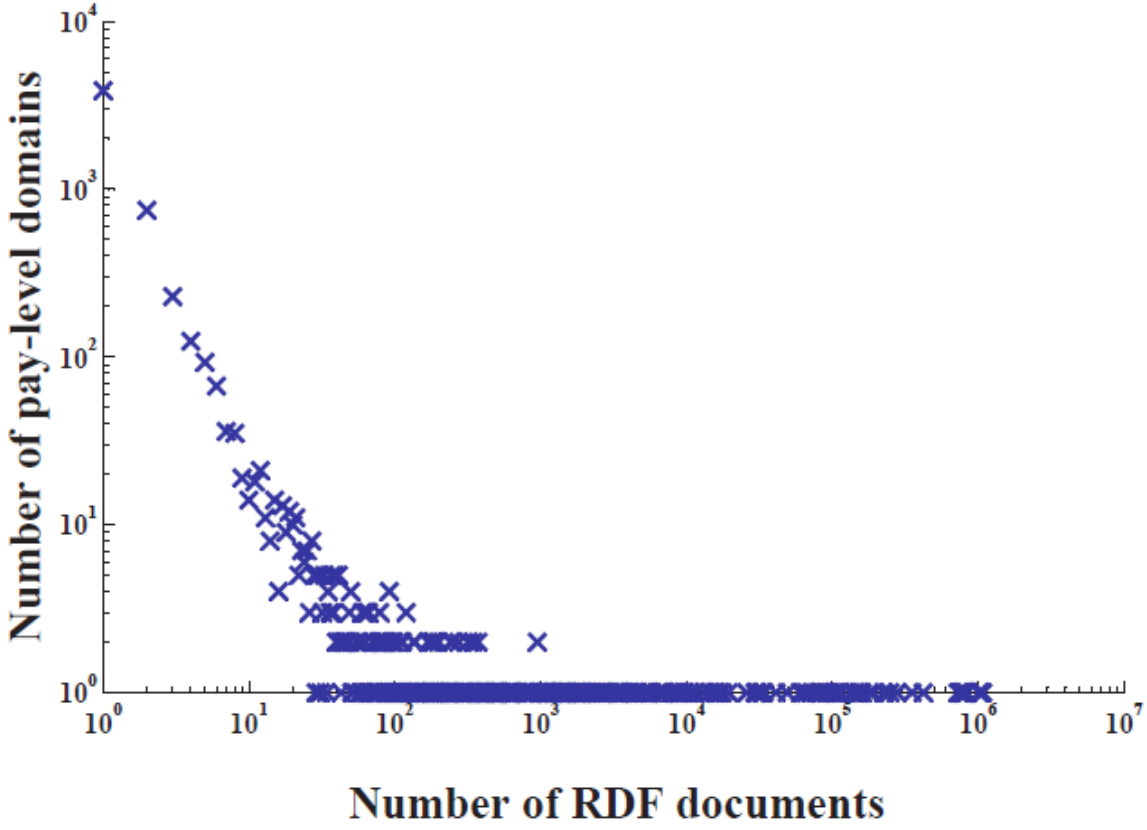
- Data set
- Vocabulary relatedness
- Post-selection vocabulary recommendation
- Conclusions

Data set statistics

- Crawled from February 2010 to May 2011 by Falcons

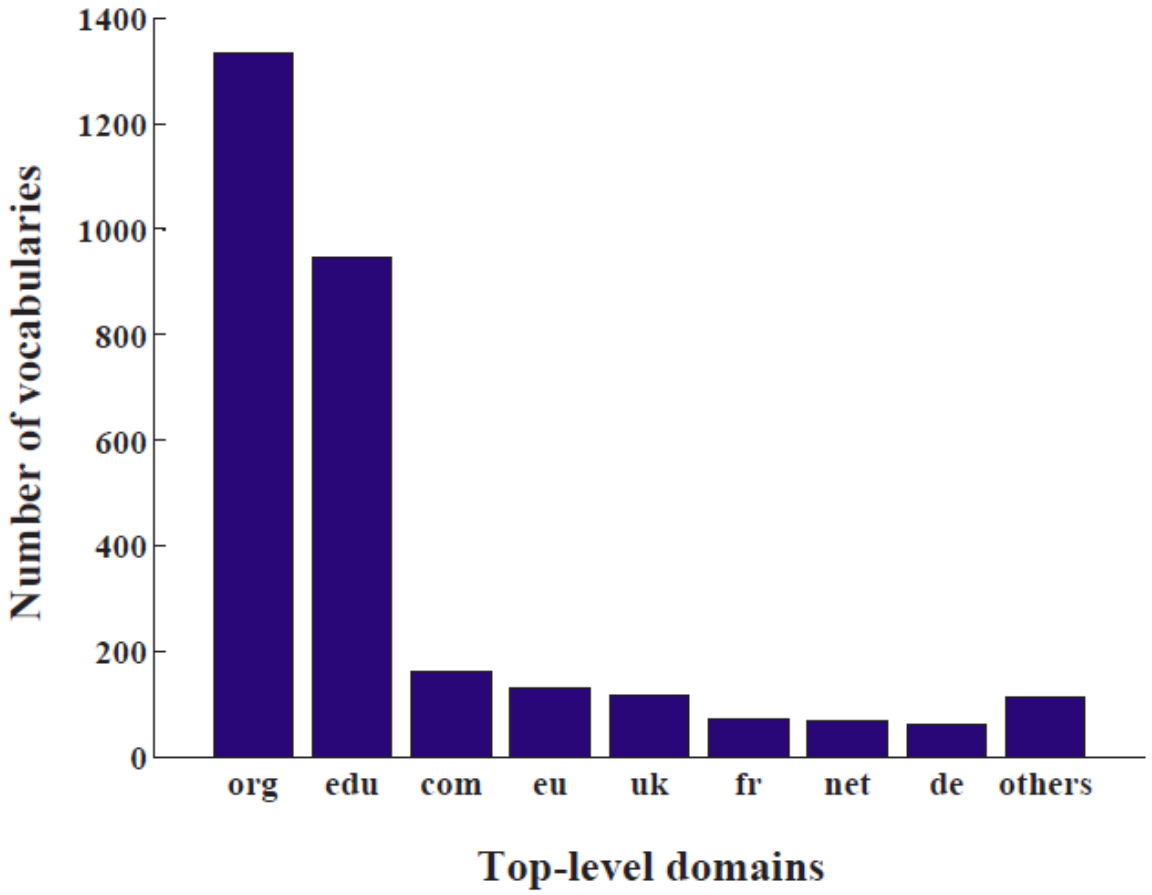
Number of RDF documents	15,947,721
Number of pay-level domains hosting RDF documents	5,805
Aggregate number of RDF triples	4,099,414,887
Number of vocabularies	2,996
Number of pay-level domains hosting vocabularies	261
Aggregate number of classes	396,023
Aggregate number of properties	59,868

- RDF documents over pay-level domains



Data set distributions

- Vocabularies over top-level domains



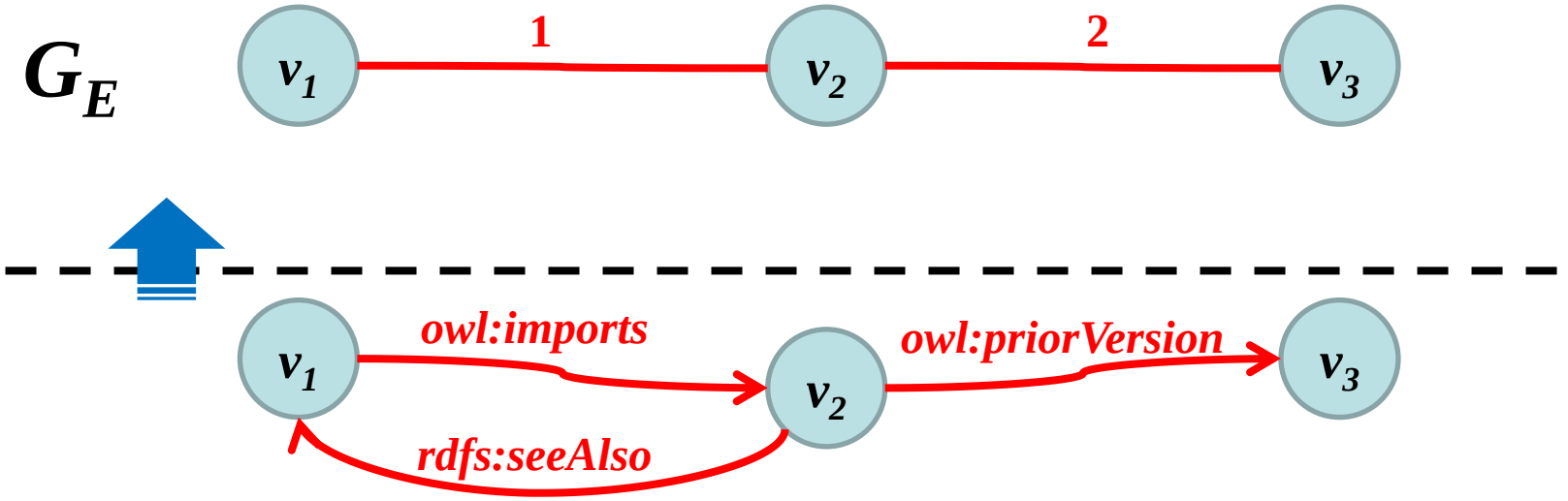
Outline

- Data set
- Vocabulary relatedness
- Post-selection vocabulary recommendation
- Conclusions

Vocabulary relatedness

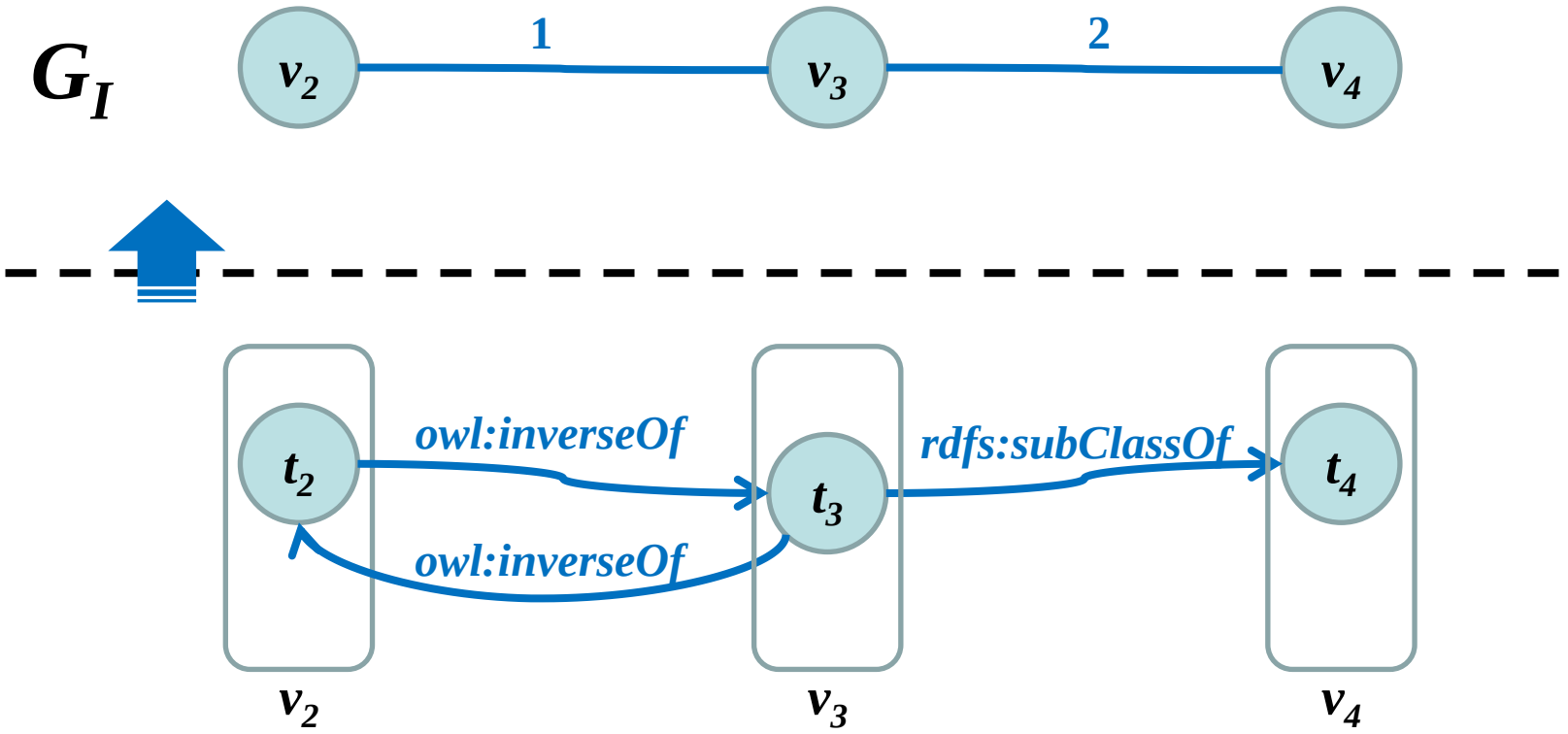
- 6 numerical measures, from 4 aspects
  - Semantic relatedness
    - Explicit
    - Implicit
    - Hybrid
  - Content similarity
  - Expressivity closeness
  - Distributional relatedness
- Comparison

- $R_S^E(v_i, v_j) = \frac{1}{\text{weight of a shortest path between } v_i \text{ and } v_j \text{ in } G_E}$



Measure 2: implicit semantic relatedness

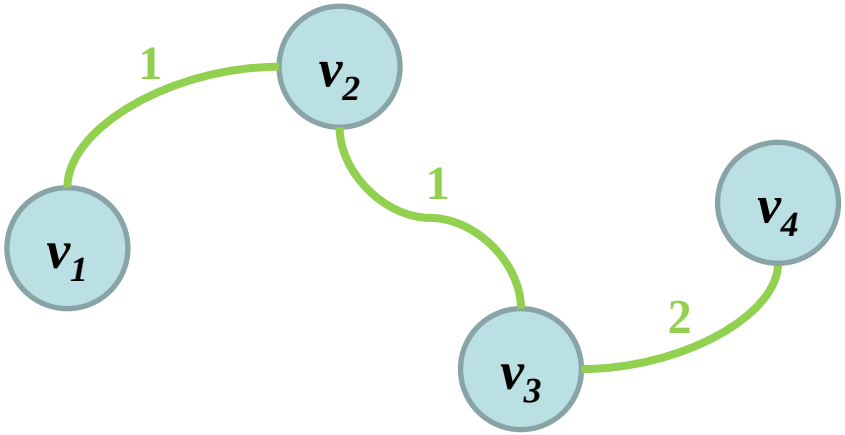
- $$R_S^I(v_i, v_j) = \frac{1}{\text{weight of a shortest path between } v_i \text{ and } v_j \text{ in } G_I}$$



Measure 3: hybrid semantic relatedness

- $$R_S^{E+I}(v_i, v_j) = \frac{1}{\text{weight of a shortest path between } v_i \text{ and } v_j \text{ in } G_{E+I}}$$

G_{E+I}



Empirical analysis (1)

- Statistical properties of G_E, G_I and G_{E+I}

	G_E	G_I	G_{E+I}
Number of nodes	2,996	2,996	2,996
Number of edges	2,968	2,845	4,691
Average degree	1.98	1.90	3.13
Maximum degree	786	684	848
Percentage of isolated nodes	56.88%	36.72%	32.31%
Number of connected components	1,763	1,143	1,007
Percentage of nodes in the largest connected component	32.78%	57.44%	62.18%
Percentage of pairs of connected nodes	5.40%	16.50%	19.33%

Empirical analysis (2)

- Explicit relations between vocabularies

http://www.w3.org/2002/07/owl#imports	36.58%
http://www.daml.org/2001/03/daml+oil#imports	1.60%
http://www.w3.org/2000/01/rdf-schema#seeAlso	0.30%
http://www.w3.org/2002/07/owl#priorVersion	0.10%
http://purl.org/dc/terms/requires	0.07%
http://www.openlinksw.com/schema/attribution#isDescribedUsing	0.07%

Measure 4: content similarity

- $$R_C(v_i, v_j) = \begin{cases} \frac{\text{SetSim}(C_i, C_j) + \text{SetSim}(P_i, P_j)}{2} & \text{if } C_i \times C_j \neq \emptyset \text{ and } P_i \times P_j \neq \emptyset \\ \text{SetSim}(C_i, C_j) & \text{if } C_i \times C_j \neq \emptyset \text{ and } P_i \times P_j = \emptyset \\ \text{SetSim}(P_i, P_j) & \text{if } C_i \times C_j = \emptyset \text{ and } P_i \times P_j \neq \emptyset \\ 0 & \text{if } C_i \times C_j = \emptyset \text{ and } P_i \times P_j = \emptyset \end{cases}$$

$$\text{SetSim}(T_i, T_j) = \text{HMean}\left(\frac{1}{|T_i|}\sum_{t_i \in T_i} \max_{t_j \in T_j} \text{LS}(t_i, t_j), \frac{1}{|T_j|}\sum_{t_j \in T_j} \max_{t_i \in T_i} \text{LS}(t_i, t_j)\right)$$

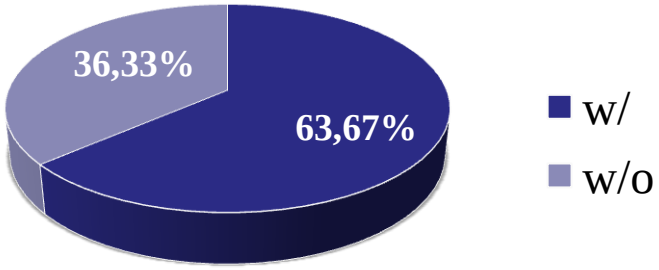
Harmonic mean

Maximum similarity between their labels

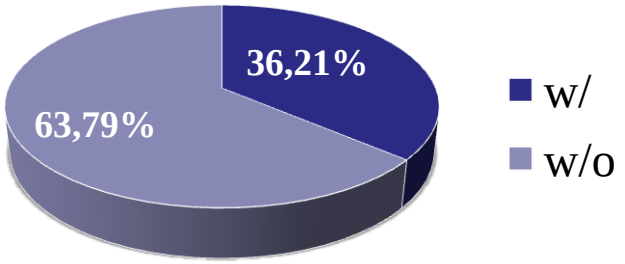
Empirical analysis (3)

- 86 label-like properties
  - rdfs:label, dc:title, and their subproperties (e.g. skos:prefLabel)
- and local name

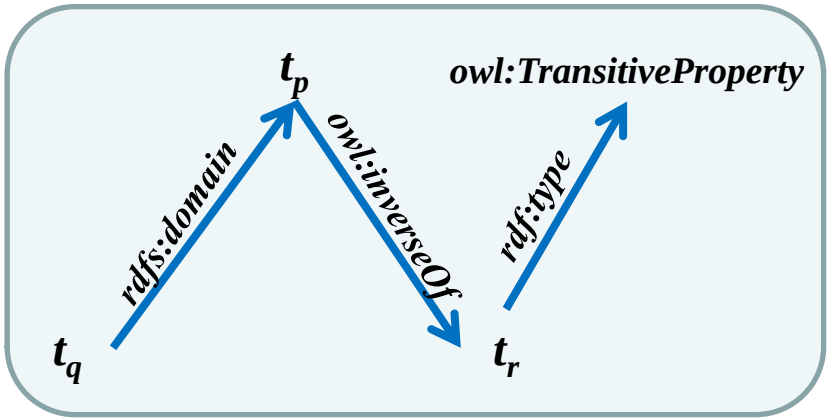
Terms and their labels

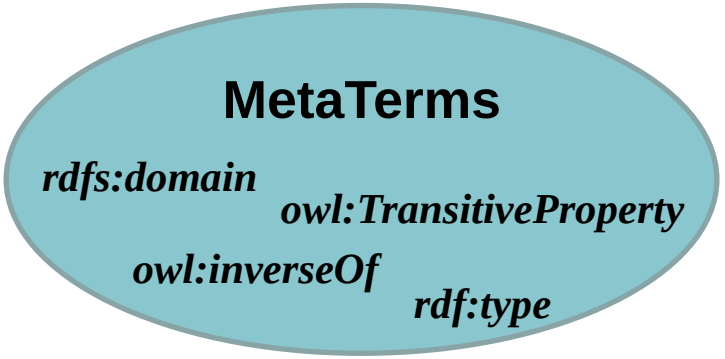


Vocabulary distribution



Measure 5: expressivity closeness





- $R_E(v_i, v_j) = \mathrm{J}(\mathrm{MetaTerms}(v_i), \mathrm{MetaTerms}(v_j))$

Jaccard

Empirical analysis (4)

- 4,978 meta-level terms, 469 (9.42%) in >1 vocabulary
- Most popular meta-level terms
  1. rdf:type
  2. rdfs:domain
  3. rdfs:range
  4. …
- and after excluding language constructs

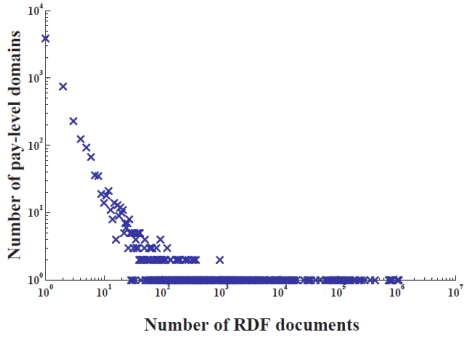
http://purl.org/dc/elements/1.1/description	1.50%
http://purl.uniprot.org/core/encodedIn	0.90%
http://www.w3.org/2004/02/skos/core#definition	0.73%
http://purl.org/dc/terms/modified	0.67%
http://www.swop-project.eu/ontologies/pmo/product.owl#unit	0.67%
http://purl.org/dc/terms/issued	0.63%
http://www.w3.org/2003/06/sw-vocab-status/ns#term_status	0.63%

- 10.13 meta-level terms per vocabulary
- ≤20 meta-level terms in 92.96% vocabularies
- but hundreds in Cyc

- Distributional profile

$$\mathrm{DP}(v) = \begin{bmatrix} p(v_1 \mid v) \\ p(v_2 \mid v) \\ \dots \\ p(v_n \mid v) \end{bmatrix} \Rightarrow R_D(v_i, v_j) = \cos(\mathrm{DP}(v_i), \mathrm{DP}(v_j))$$

$$\mathrm{DP}_i(v) = \frac{|\{d \in D \mid v, v_i \in \mathrm{IV}(d)\}|}{|\{d \in D \mid v \in \mathrm{IV}(d)\}|}$$



$$\mathrm{DP}_i(v) = \frac{|\{S \in \mathrm{PLD}(D) \mid \exists d \in S, v, v_i \in \mathrm{IV}(d)\}|}{|\{S \in \mathrm{PLD}(D) \mid \exists d \in S, v \in \mathrm{IV}(d)\}|}$$

Empirical analysis (5)

- Instantiation found for 1,874 (62.55%) vocabularies
- Most popular vocabularies (excluding languages)

http://purl.org/dc/elements/1.1/	37.45%
http://xmlns.com/foaf/0.1/	22.79%
http://purl.org/dc/terms/	15.90%
http://www.icra.org/rdfs/vocabularyv03#	10.65%
http://www.w3.org/2003/01/geo/wgs84_pos#	5.22%
http://purl.org/vocab/bio/0.1/	2.76%
http://www.w3.org/2000/10/swap/pim/contact#	2.76%
http://rdfs.org/sioc/ns#	2.20%
http://usefulinc.com/ns/doap#	1.67%
http://purl.org/vocab/relationship/	1.38%

Empirical analysis (6)

- Co-instantiation found for 9,763 pairs of vocabularies
- Most popular vocabulary co-instantiation (excluding languages)

http://purl.org/dc/elements/1.1/ http://purl.org/dc/terms/	14.42%
http://purl.org/dc/elements/1.1/ http://www.icra.org/rdfs/vocabularyv03#	10.65%
http://purl.org/dc/terms/ http://www.icra.org/rdfs/vocabularyv03#	10.61%
http://xmlns.com/foaf/0.1/ http://purl.org/dc/elements/1.1/	9.42%
http://www.w3.org/2003/01/geo/wgs84_pos# http://xmlns.com/foaf/0.1/	5.05%

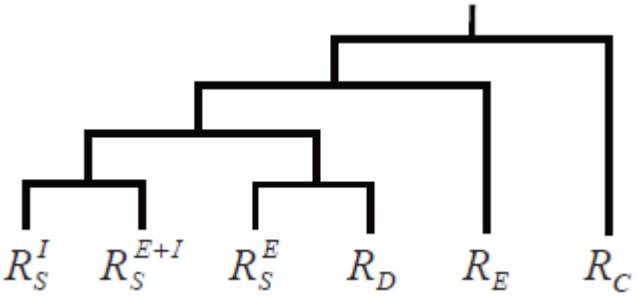
Vocabulary relatedness

- 6 numerical measures, from 4 aspects
  - Semantic relatedness
    - Explicit
    - Implicit
    - Hybrid
  - Content similarity
  - Expressivity closeness
  - Distributional relatedness
- Comparison

- Spearman's rank correlation coefficient (ρ∈[-1,1])

	R_S^I	R_S^{E+I}	R_C	R_E	R_D
R_S^E	0.39	0.53	0.21	0.19	0.66
R_S^I	-	0.88	0.26	0.38	0.35
R_S^{E+I}	-	-	0.30	0.26	0.43
R_C	-	-	-	0.32	0.23
R_E	-	-	-	-	0.24

- Single-link hierarchical clustering



Outline

- Data set
- Vocabulary relatedness
- Post-selection vocabulary recommendation
- Conclusions

Relatedness-based ranking

- Ranking by single measure: $R_j(v_i, v_0)$

  $R_j \in \mathfrak{R}$

  $\mathfrak{R} = \{R_S^E, R_S^I, R_S^{E+I}, R_C, R_E, R_D\}$

- Ranking by multiple measures: $\sum_{R_j \in \mathfrak{R}} \alpha_j R_j(v_i, v_0)$

Popularity-based re-ranking

- $$\sum_{R_j \in \mathfrak{R}} \alpha_j R_j(v_i, v_0) \cdot (1 + \log_b (1 + \mathrm{Pop}(v_i)))$$

Degree of influence of popularity

Number of pay-level domains instantiating v_i

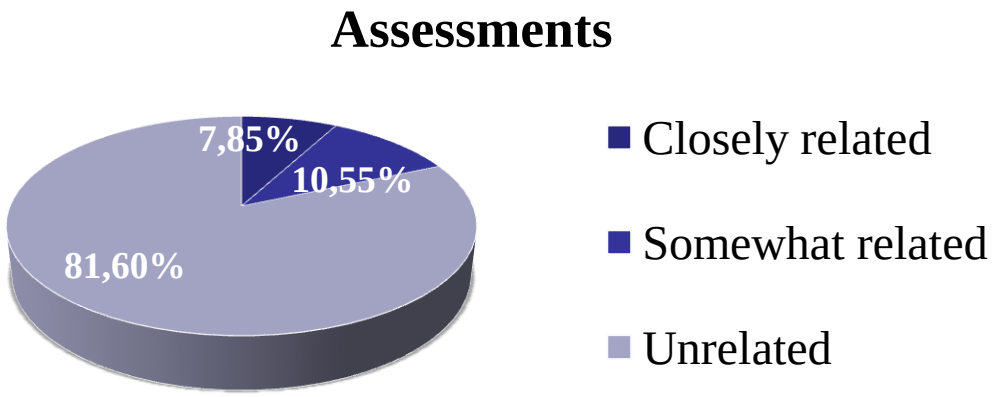
Evaluation settings

- 20 “selections” randomly selected from 1,302 moderate-sized vocabularies
- Depth-10 pooling with $\mathfrak{R} = \{R_S^E, R_S^I, R_S^{E+I}, R_C, R_E, R_D\}$

- 2 experts
- Ratings
  - Closely related: 2
  - Somewhat related: 1
  - Unrelated: 0

- Metric: NDCG

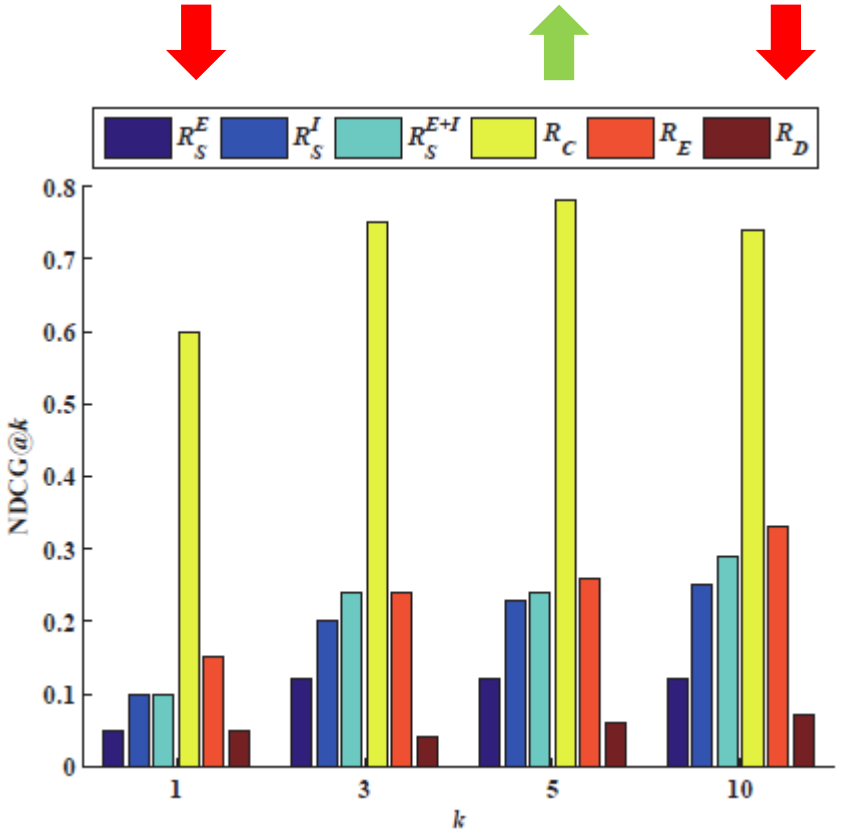
- 739 assessments



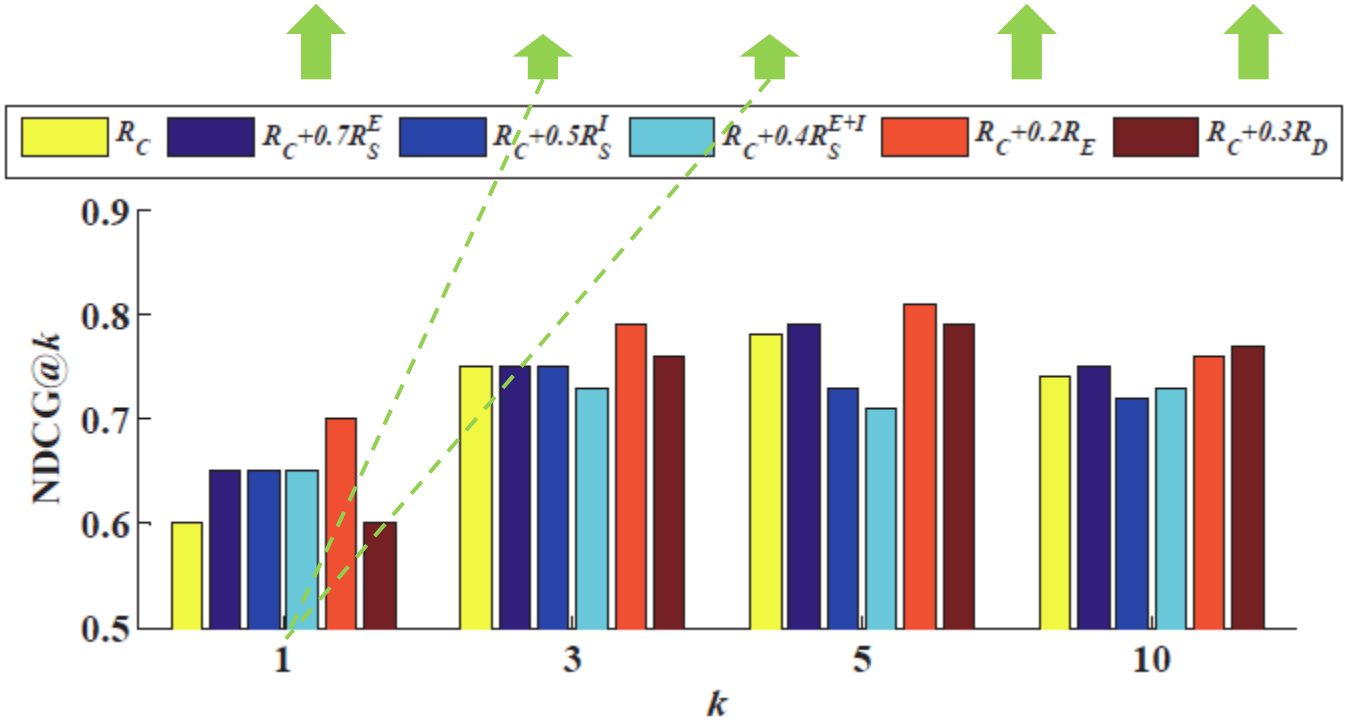
- Agreement between experts
  - 80%
  - or 91% when “closely related = somewhat related = related”

56.88% isolated vocabularies in G_E

37.45% uninstantiated vocabularies

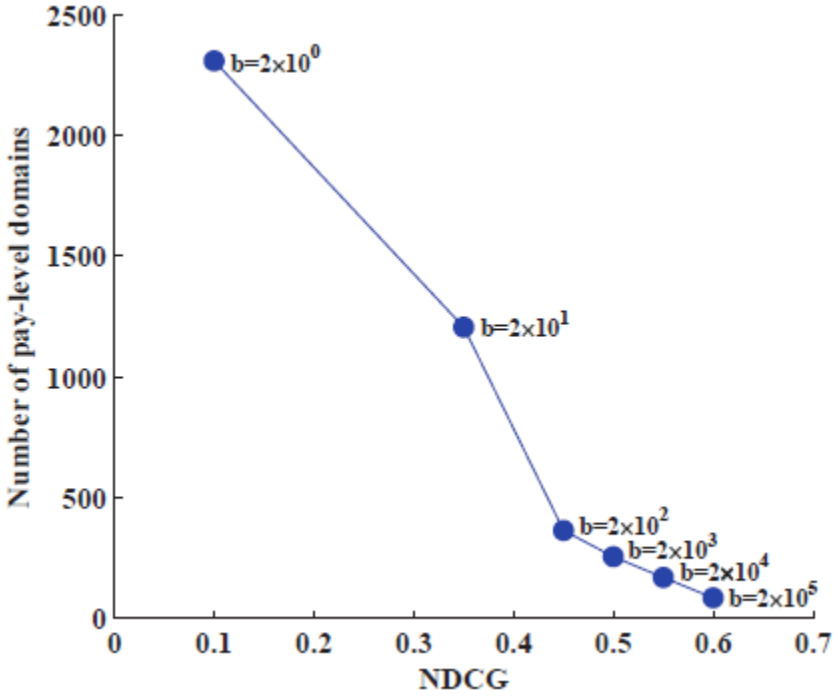


Evaluation results --- combinations of measures



Relatedness vs. popularity

- NDCG@1 vs. number of pay-level domains instantiating it



Outline

- Data set
- Vocabulary relatedness
- Post-selection vocabulary recommendation
- Conclusions

Conclusions

- Vocabulary-level relatedness
  - 4 aspects, 6 measures
- Empirical analysis
  - Statistical findings
  - Comparison
- Post-selection vocabulary recommendation
  - Relatedness-based ranking
  - Popularity-based re-ranking
  - Evaluation

- Falcons Ontology Search
  - http://ws.nju.edu.cn/falcons/ontologysearch/

Take away

- Vocabulary meta-descriptions are incomplete.
- Terms lack labels.
- Co-instantiated ∝ explicitly related

Falcons

http://ws.nju.edu.cn/falcons/ontologysearch/