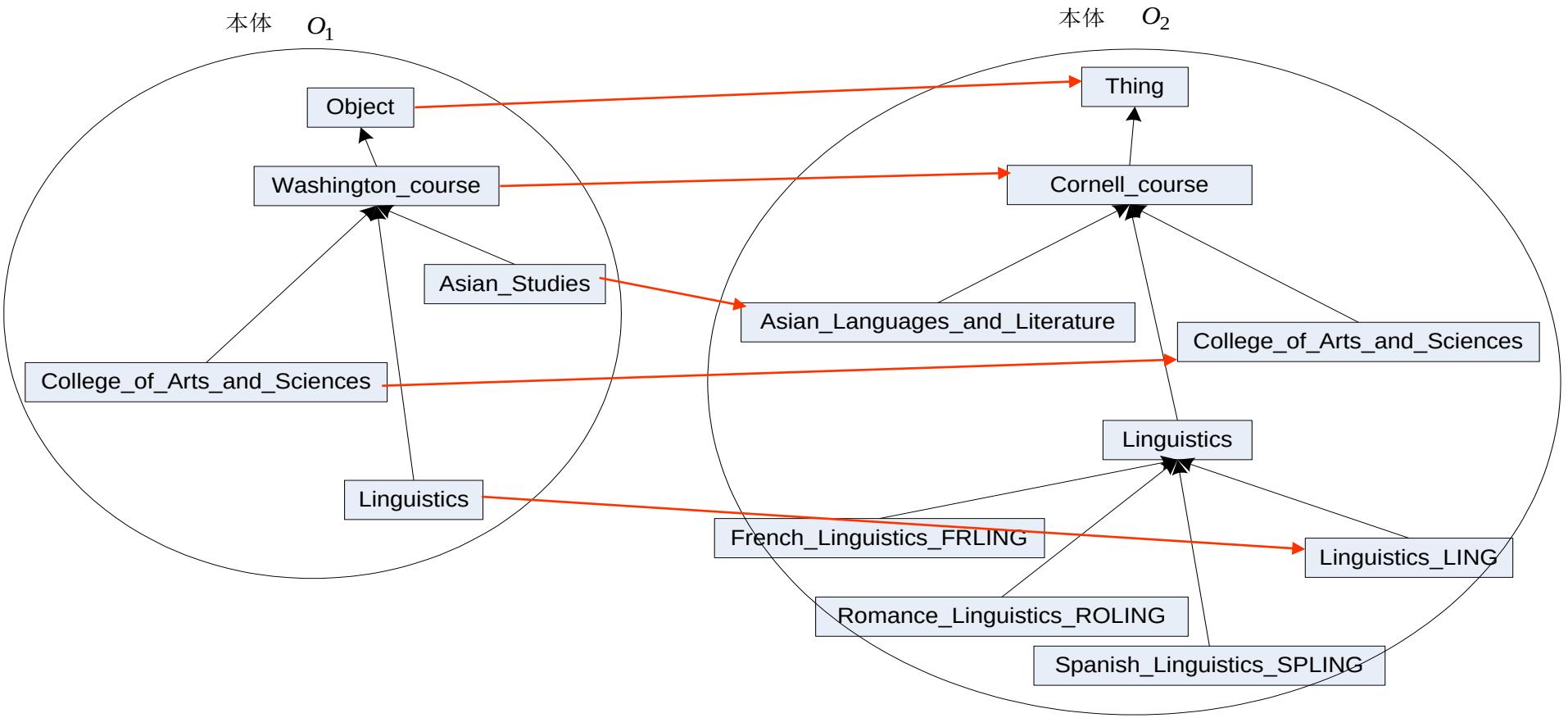
From Dynamic to Unbalanced Ontology Matching

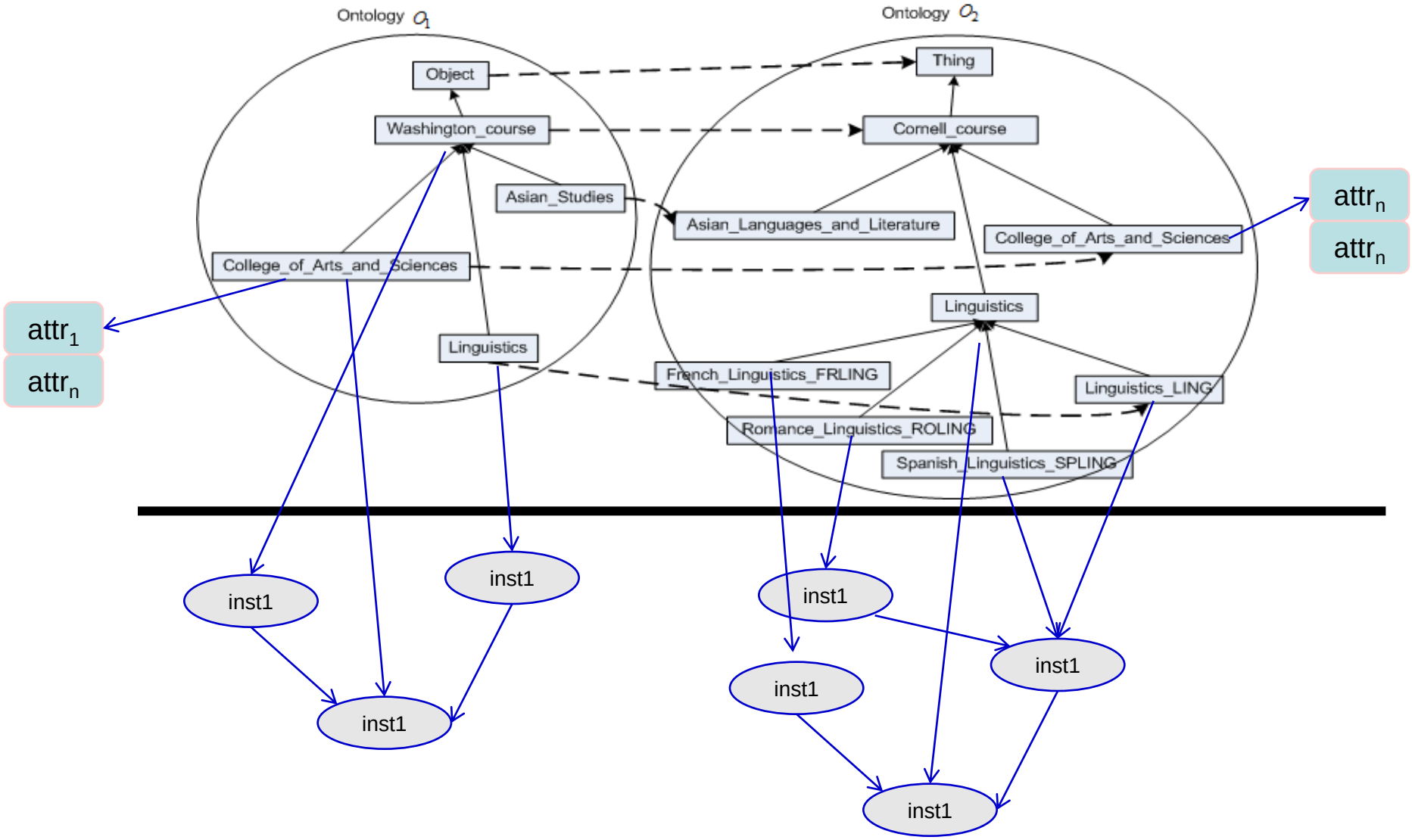
Jie Tang

Knowledge Engineering Group,
Dept. of Computer Science and Technology
Tsinghua University
May 22th 2009

What is Ontology Matching?



Ontology Matching

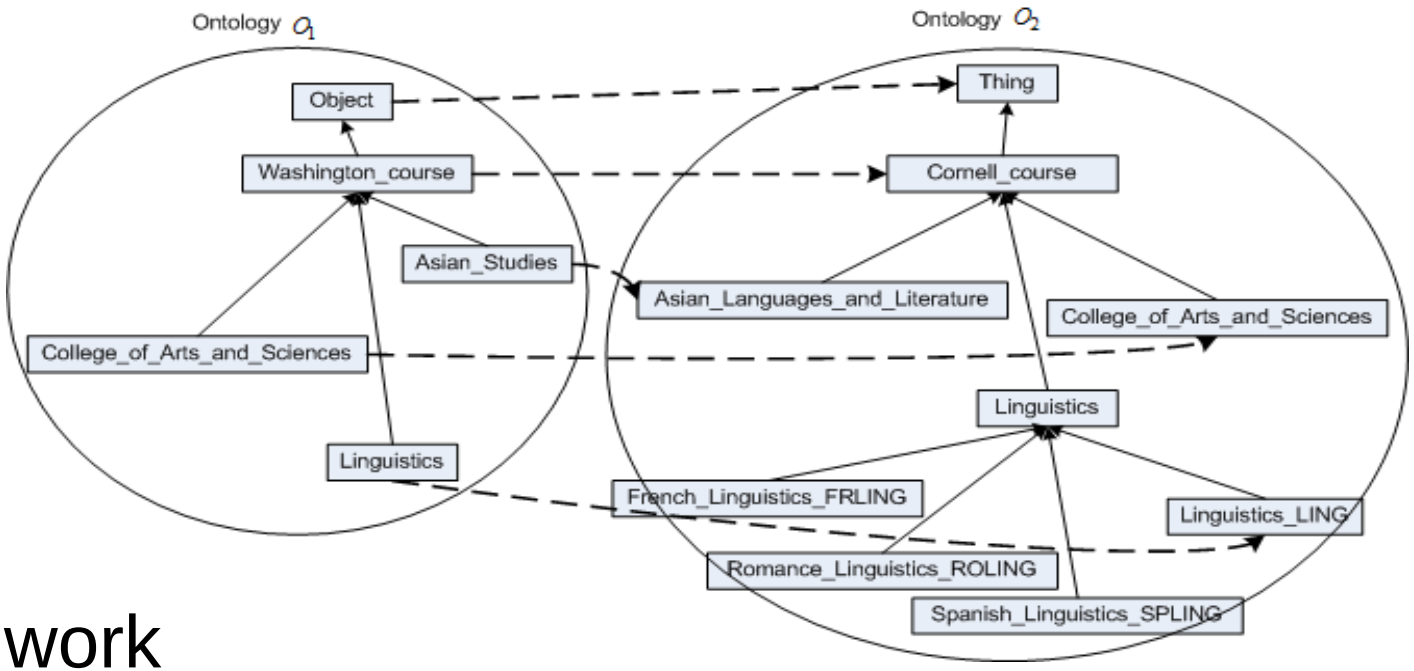


Problem Definition

Matching Function: $Map(\{e_{i1}\}, O_1, O_2) = \{e_{i2}\}$

Cardinality	O_1	O_2	**Mapping Expression**
1:1	Faculty	Academic staff	O_1.Faculty= O_2.Academic staff
1:n	Name	First name, Last name	O_1.Name= O_2.First name+O_2.Last name
n:1	Cost, Tax ratio	Price	O_1.Cost*(1+ O_1.Tax ratio)= O_2.Price
1:null	AI		
null:1		AI	
n:m	BookTitle, BookaNo, PublisherNo, PublisherName	Book, Publisher	O_1.BookTitle + O_1.BookaNo + O_1.PublisherNo + O_1.PublisherName = O_2.Book + O_2.Publisher

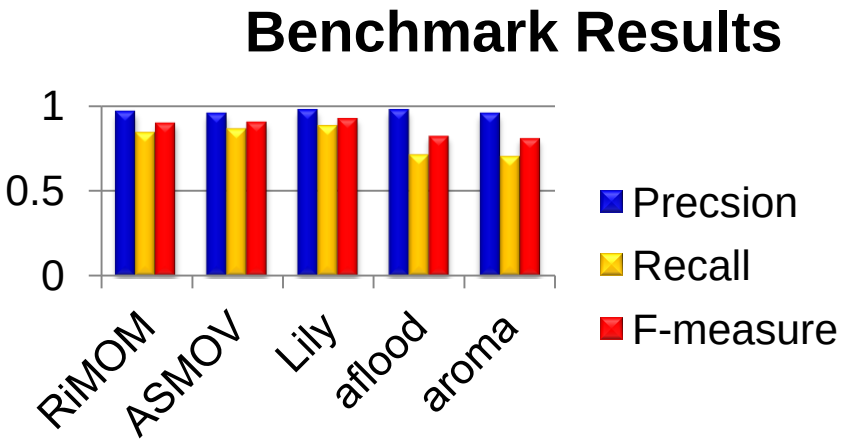
Ontology Matching

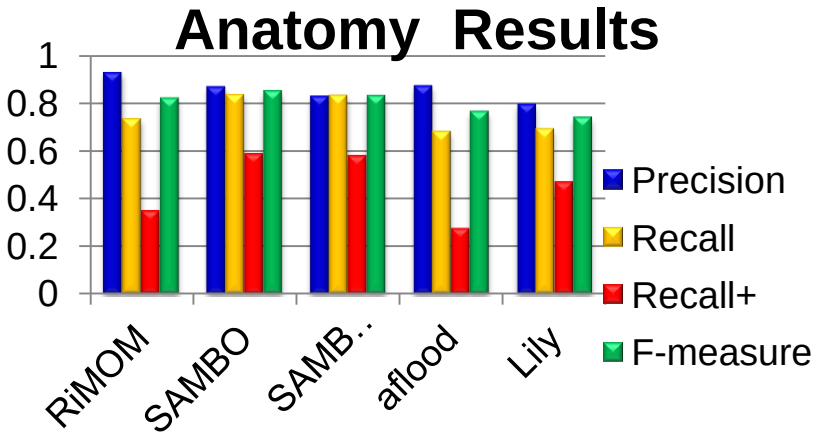


- Our work
  - RiMOM: Risk minimization based approach
    - Jie Tang, et al. Journal of Web Semantics. 2006, **Dec**. (**JoWS**, **IF： 3.41**)
  - Dynamic ontology matching framework
    - Juanzi Li, Jie Tang, et al. **TKDE**, 2009
  - Unbalanced ontology matching
    - Qian Zhong, Hanyu Li, Juanzi Li, Guotong Xie, Jie Tang. **SIGMOD'2009**.

RiMOM—A tool for ontology matching

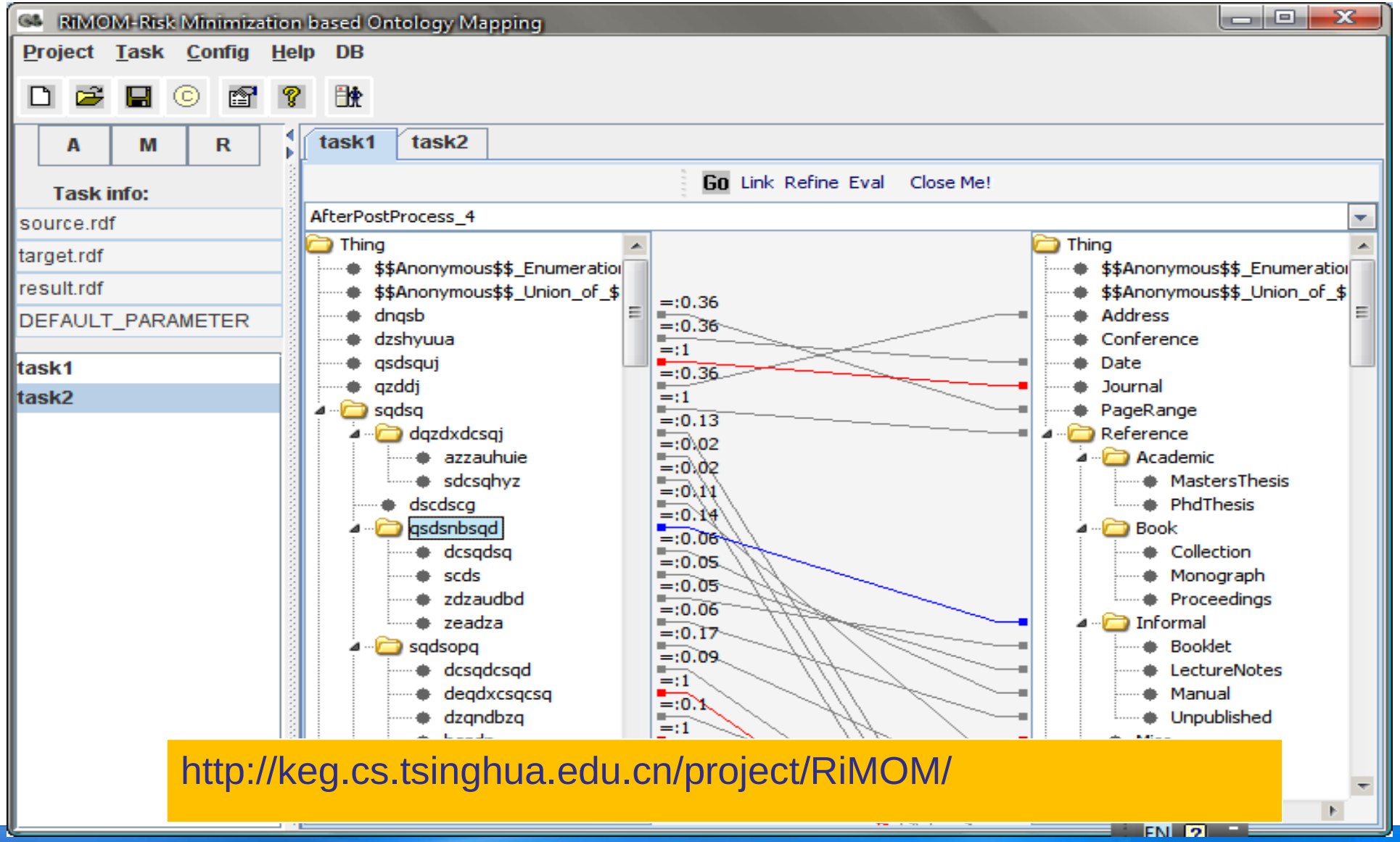
- OAEI(2006-2008):an international contest on ontology alignment





Msg from Chair: "I'm really surprised by the good results of these years RiMOM, you can compete with the top systems that make use of such background knowledge."

RiMOM—A tool for ontology matching



http://keg.cs.tsinghua.edu.cn/project/RiMOM/

Outline

- Dynamic Multi-strategy Ontology Matching
- Unbalanced Ontology Matching
- Discussion

A Dynamic Multi-strategy Ontology Alignment Framework

- Matching = Multi-strategies + Strategy selection

- Concept/Attribut
- Concept/Attribut
- Concept/Attribut
- Instance
- Structure

- Linguistic similarity factor

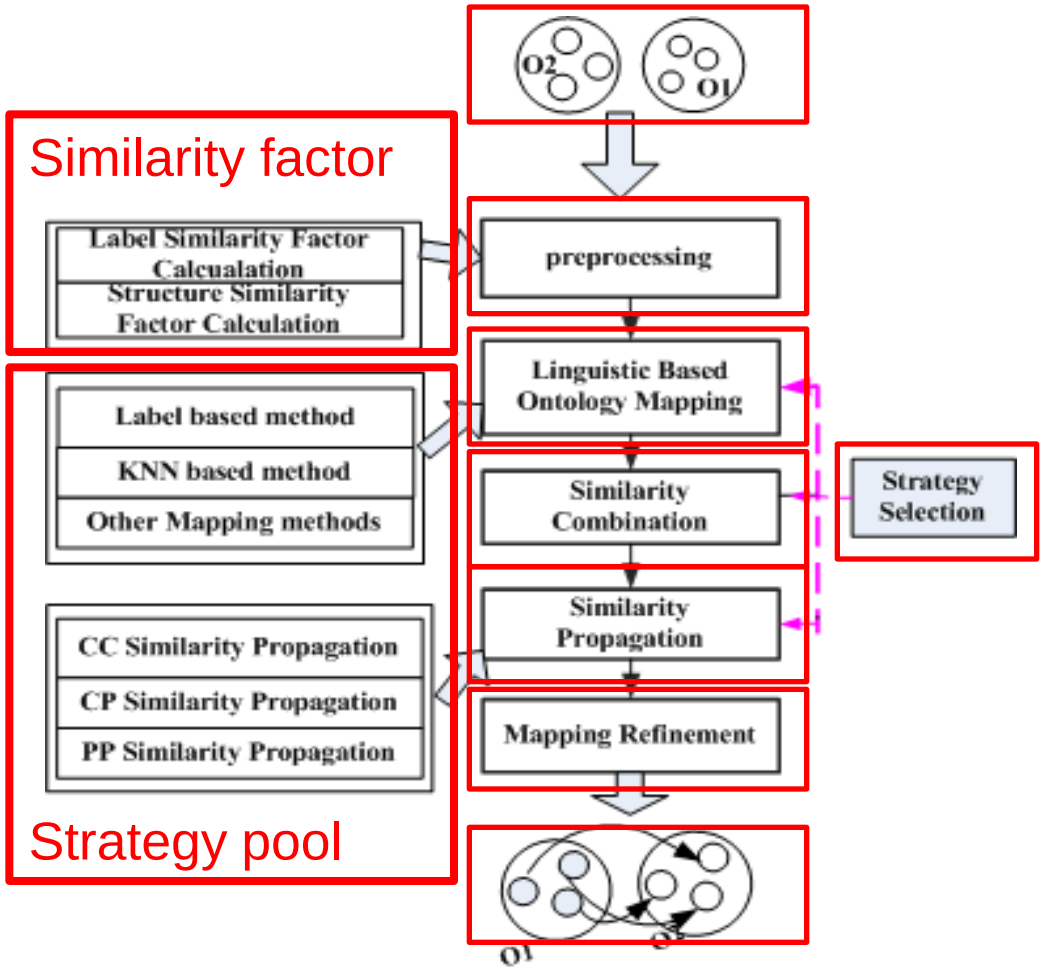
$$F_LS = \frac{\#same_label}{\max(\#c_1, \#c_2)}$$

- Structural similarity factor

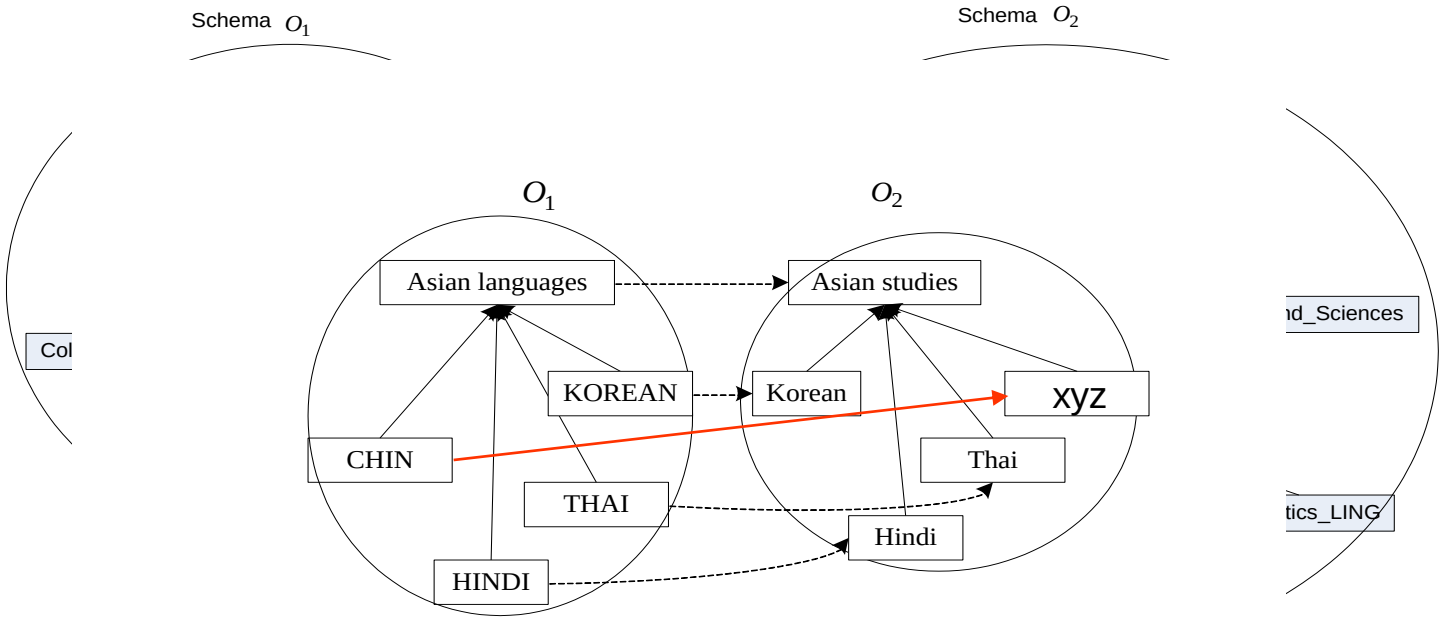
$$F_SS = \frac{\#common_concept}{\max(\#nonleaf_c_1, \#nonleaf_c_2)}$$

- Associate a loss for each candidate matching
- Strategy selection: determine if we should use the strategies

A General Processing Flow



Multiple Strategies



- Concept name: similarity(washington_course, cornell_course)
- Concept path: similarity(/object/washington_course, /thing/cornell_course)
- Concept description: classifier = train(O_2) and classify (O_1, classifier)
- Instance: classifier = train(O_2) and classify (O_1, classifier)
- Structure: taxonomy information. E.g. Hypernyms and Hyponyms

Multiple Linguistic Strategies

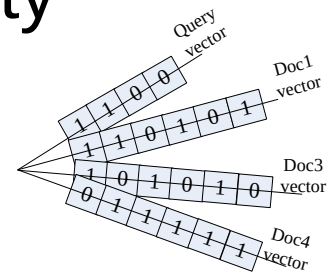
- Edit distance on entity's label

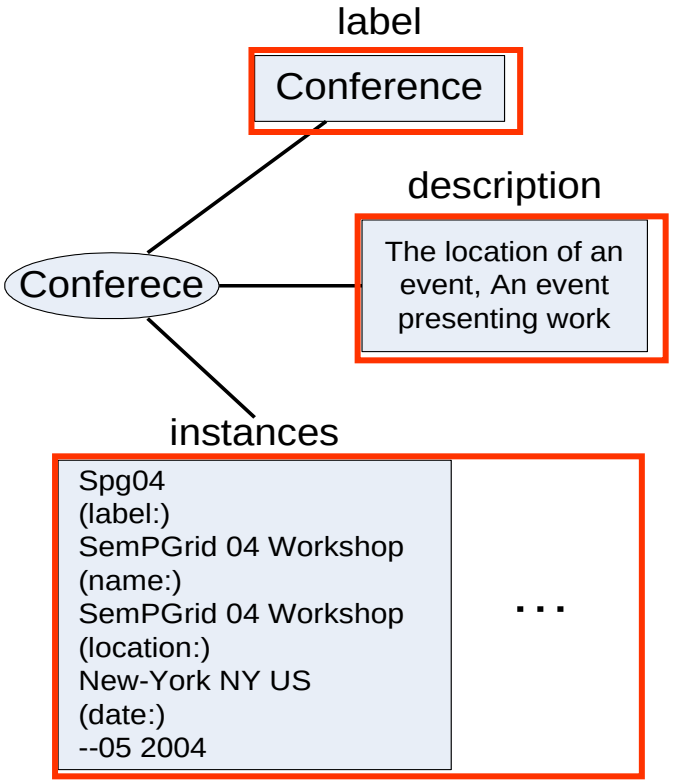
$$s_{edit} = \frac{1}{1 + edit_distance(w_i, w_j)}$$

$$edit_distance(w_i, w_j) = \frac{|\{op_k\}|}{\max(length(w_i), length(w_j))}$$

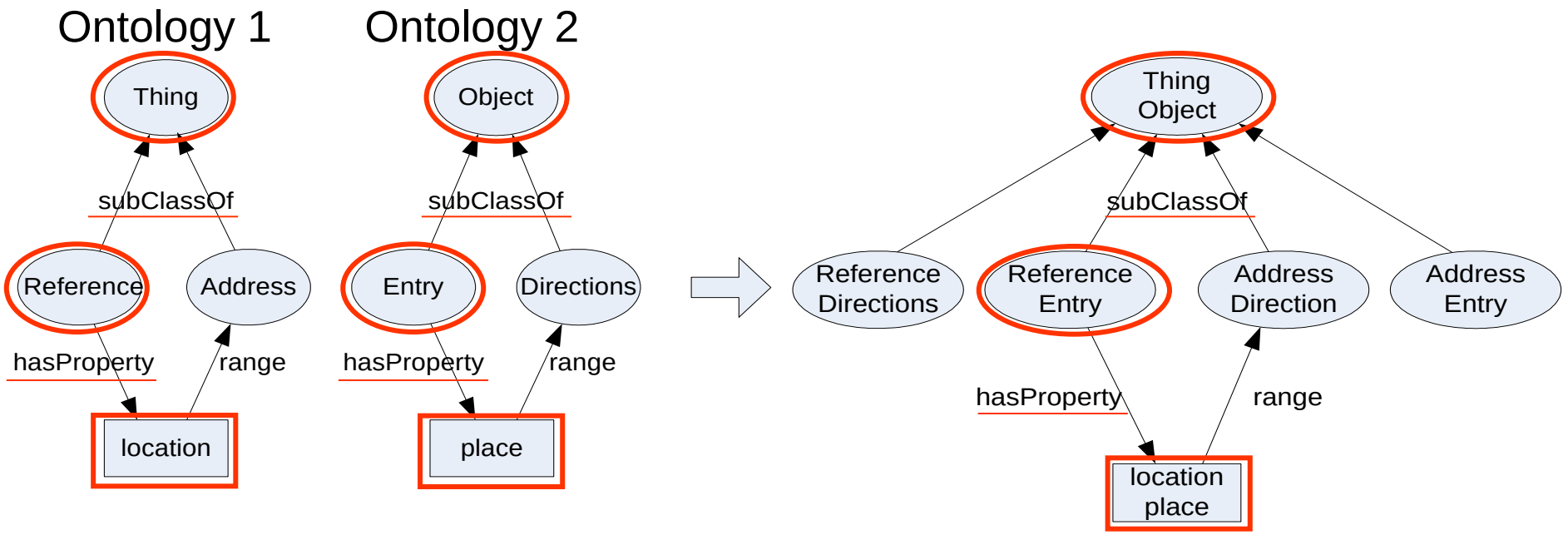
- WordNet: $s_{wordnet}(w_i, w_j) = \frac{2 \times \log p(s)}{\log p(s_i) + \log p(s_j)}$

- Vector-based similarity





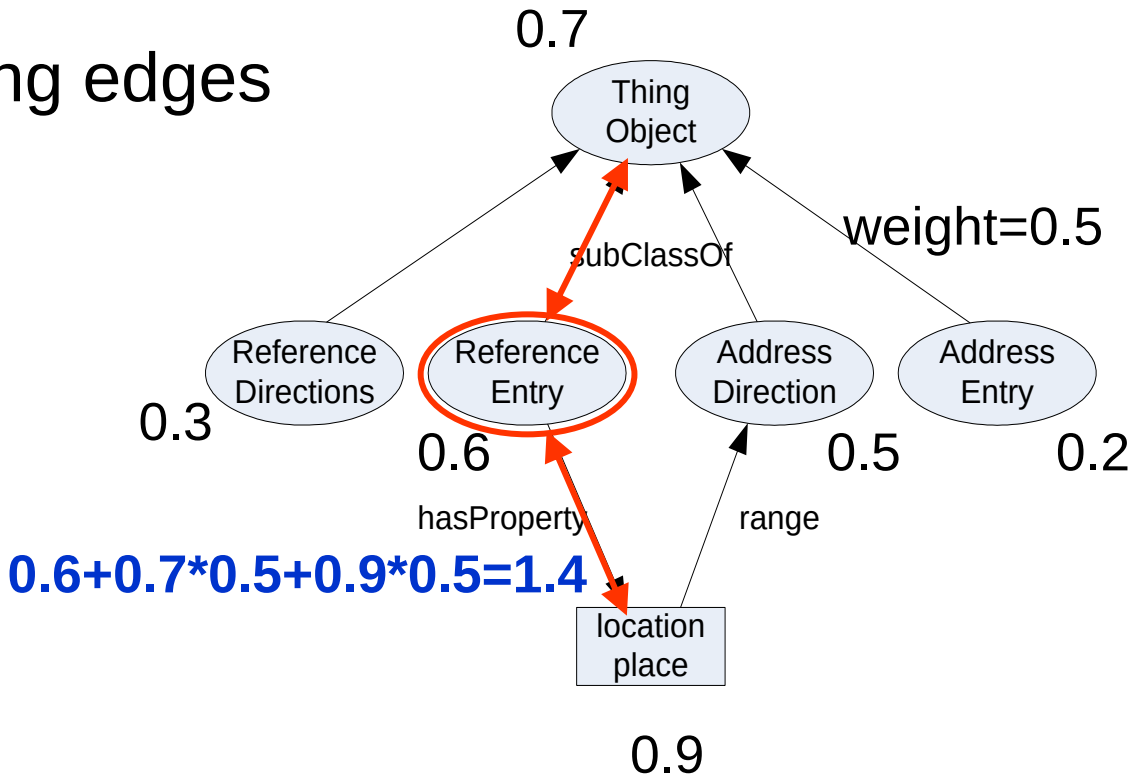
Similarity Propagation



The construction of an intermediate graph from original ontologies

Similarity Propagation (cont.)

- Propagate similarities along edges
- Three types of edges:
  - Class to Class (CCP)
  - Class to Property (CPP)
  - Property to Property (PPP)



Strategy Pool

Strategy pool

Edit-distance

Sim = 1-ED($label_1$, $label_2$)

Vector-similarity

- weighted vector generation
  - content feature
  - structure feature
- cosine similarity

Background-knowledge

- external knowledge
- similarity definition

Similarity-combination

$$Map(e_1, e_2) = \frac{\sum_{k=1...n} w_k \sigma(Map_k(e_1, e_2))}{\sum_{k=1...n} w_k}$$

Path-similarity

- entity path
- path similarity definition

Similarity-propagation

- three propagation strategies
- CCP, PPP, CPP

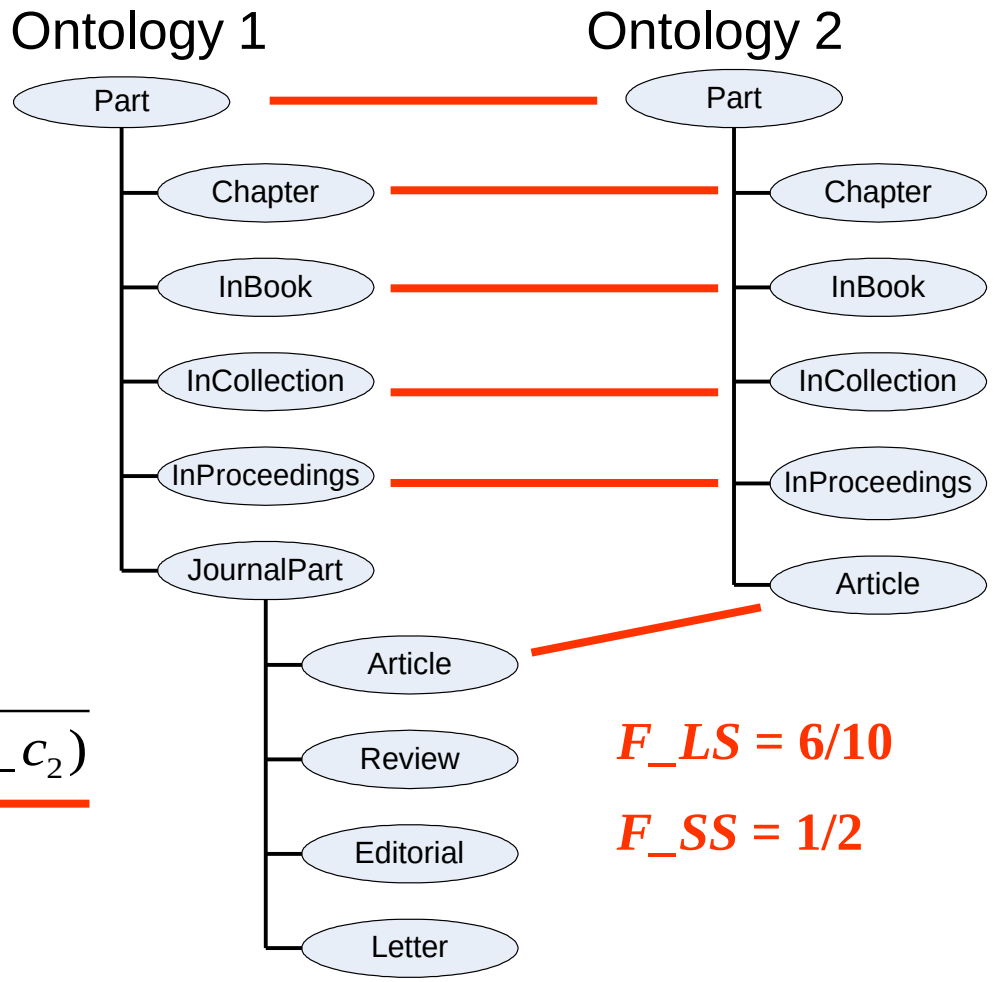
Strategy Selection—Similarity factor

- Label similarity factor

$$F_LS = \frac{\#same_label}{\max(\#c_1, \#c_2)}$$

- Structure similarity factor

$$F_SS = \frac{\#common_concept}{\max(\#nonleaf_c_1, \#nonleaf_c_2)}$$



$\max(\#nonleaf_c_1, \#nonleaf_c_2) = 2$

Strategy Selection

- Strategy Selection
  - Selection with the two similarity factors
  - Determining whether a strategy is to be used in the alignment process
  - E.g. if $F_SS>0.25$, we use CCP, CPP, and PPP for propagation. …
- Linguistic Strategy
  - Adding structural features in vector-based similarity

Outline

- Dynamic Multi-strategy Ontology Matching
  - Experimental Results
- Unbalanced Ontology Matching
- Discussion

Data Sets

- OAEI 2006
  - Benchmark (15-69), 53 alignment tasks
  - Directory: (4,500), Yahoo and ODP
  - Food: (16,000 vs. 41,000), two SKOS thesaurus
- OAEI 2007
- Comparison methods

Notation	Description
NA+VA	Edit distance + Vector distance based strategies
RiMOM	Full-fledged RiMOM
RiMOM-SP	RiMOM without similarity flooding
RiMOM-SS	RiMOM without strategy selection

Statistics on the Data Set

Data set	Ontology	#concept	#attribute	#alignment（ground truth）	#instance
Benchmark	Reference Ontology	33	59	--	76
	101	33	61	91	111
	103	33	61	91	111
	104	33	61	91	111
	201	34	62	91	111
	202	34	62	91	111
	204	33	61	91	111
	205	34	61	91	111
	221	34	61	91	111
	222	29	61	91	111
	223	68	61	91	111
	224	33	59	91	0
	225	33	61	91	111
	228	33	0	33	55
	230	25	54	75	83
	301	15	40	61	0
	302	15	31	48	0
	303	54	72	49	0
	304	39	49	76	0

Similarity between Ontologies

Description of Benchmark Data Set

Name of test sets	Test sets	Ontology characteristics	Number of ontologies
D1	101-104	Similar both in label description and hierarchy structure	4
D2	201-210	Similar in hierarchy structure	10
D3	221-247	Similar in label description	18
D4	248-266	Different in both label description and hierarchy structure	15
D5	301-304	Real world ontologies defined by different institutions	4

Ontology	F_LS	F_SS
#104	1.00	0.90
#201	0.05	1.00
#205	0.20	1.00
#221	1.00	0.00
#222	0.95	0.83
#233	1.00	0.00
#254	0.10	0.00
#301	0.20	0.00
#302	0.61	0.33
#304	0.80	0.54

Results on OAEI2006

Test	1xx		2xx		3xx		H-mean	
System	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
Refalign	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Edna	0.96	1.00	0.94	0.67	0.94	0.61	0.91	0.54
Automs	0.94	1.00	0.94	0.89	0.91	0.70	0.94	0.67
Coma	1.00	1.00	0.99	0.97	0.84	0.69	0.96	0.83
DSSim	1.00	0.98	0.99	0.67	0.90	0.78	0.98	0.55
Falcon	1.00	1.00	0.97	0.97	0.89	0.78	0.92	0.86
Hmatch	0.91	1.00	0.84	0.70	0.78	0.57	0.84	0.55
Jhuapl	1.00	1.00	0.50	0.98	0.18	0.50	0.22	0.85
OCM	0.95	1.00	0.94	0.69	0.89	0.51	0.93	0.55
Prior	1.00	1.00	0.96	0.79	0.85	0.80	0.95	0.63
RiMOM	1.00	1.00	1.00	0.98	0.83	0.82	0.96	0.88

RiMOM vs. RiMOM-SP

Test set	RiMOM-SP			RiMOM		
	P	R	F1	P	R	F1
D1	100.0	100.0	100.0	100.0	100.0	100.0
D2	91.5	82.3	86.7	98.5	95.3	96.9
D3	98.9	99.9	99.4	98.9	100.0	99.4
D4	89.2	36.8	52.1	90.2	54.1	67.6
D5	80.1	80.9	80.5	81.5	81.0	81.2
Overall	94.5	78.6	85.8	95.9	88.1	91.8

RiMOM vs. RiMOM-SS

Test set	RiMOM-SS			RiMOM		
	P	R	F1	P	R	F1
D1	100.0	100.0	100.0	100.0	100.0	100.0
D2	96.2	92.9	94.5	98.5	95.3	96.9
D3	98.9	99.9	99.4	98.9	100.0	99.4
D4	85.8	49.7	62.9	90.2	54.1	67.6
D5	78.3	79.4	78.8	81.5	81.0	81.2
Overall	93.5	85.5	89.3	95.9	88.1	91.8

Relationship with Several Classical Methods

T - metadata textural content; S - metadata structure;

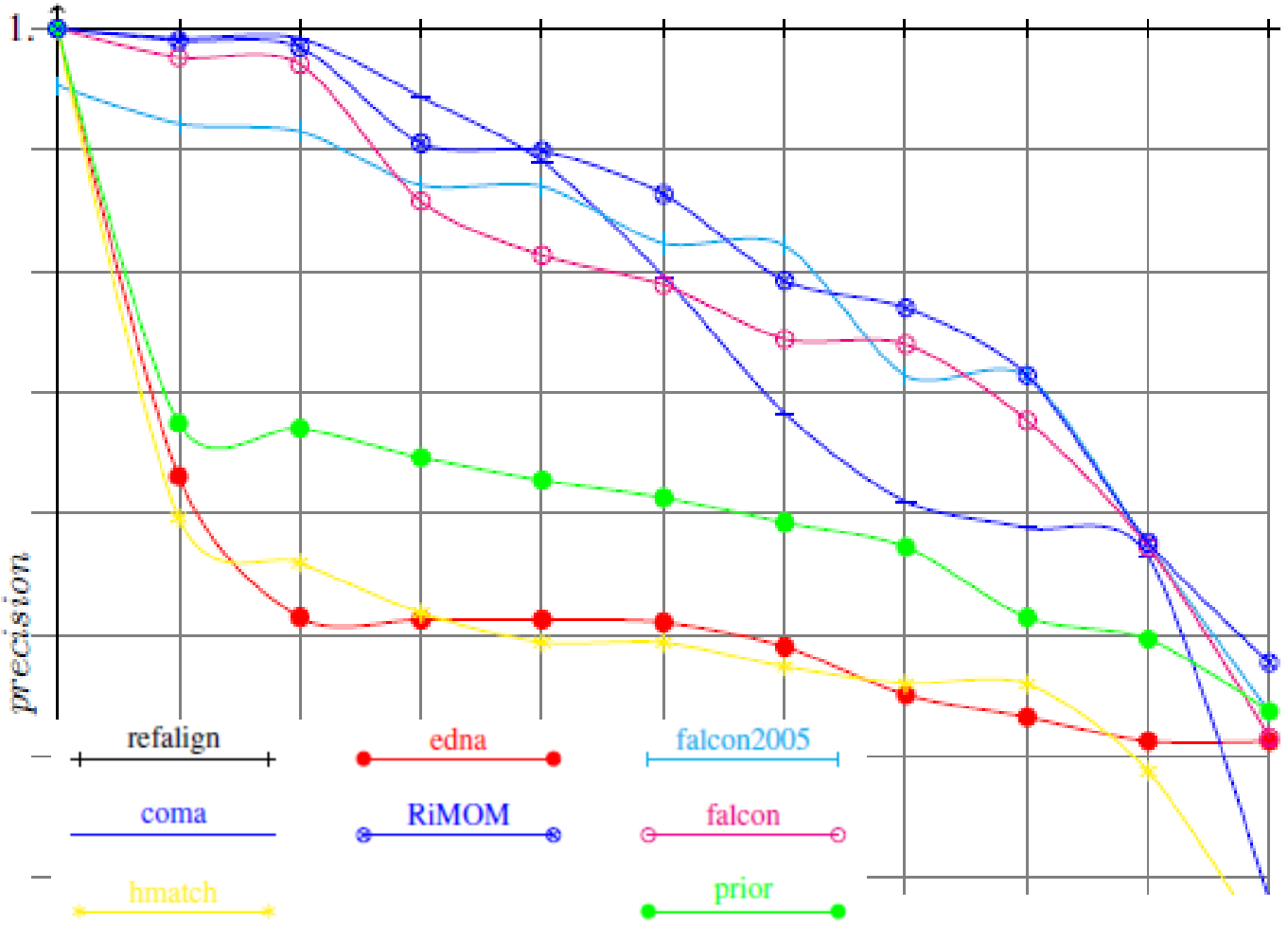
I - ontological instance; K - domain knowledge.

L - learning-based; M - similarity matching-based;

R - reasoning-based; D - Dynamic strategy

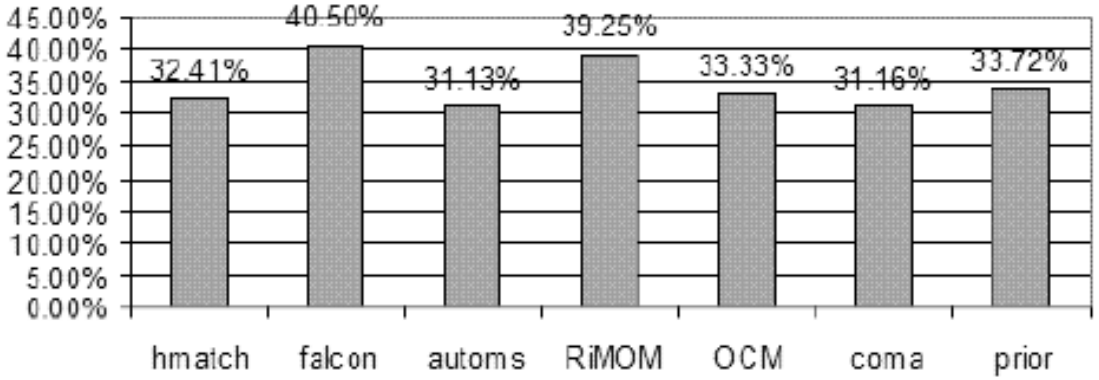
Method	Used Information				Strategy			
	T	S	I	K	L	M	R	D
Coma [2] [9]	√	√	√	√		√		
Falcon [29] [53]	√	√	√			√		
GLUE [10] [11]			√		√			
PROMPT [49]	√	√				√		
Rondo [47]	√	√		√		√		
S-Match [24]	√	√				√	√	
Uniform [52]	√	√	√			√		
Pre-RiMOM [60]	√	√	√	√	√	√		
RiMOM	√	√	√			√		√

Results on OAEI 2006



Results on OAEI2006

- Directory



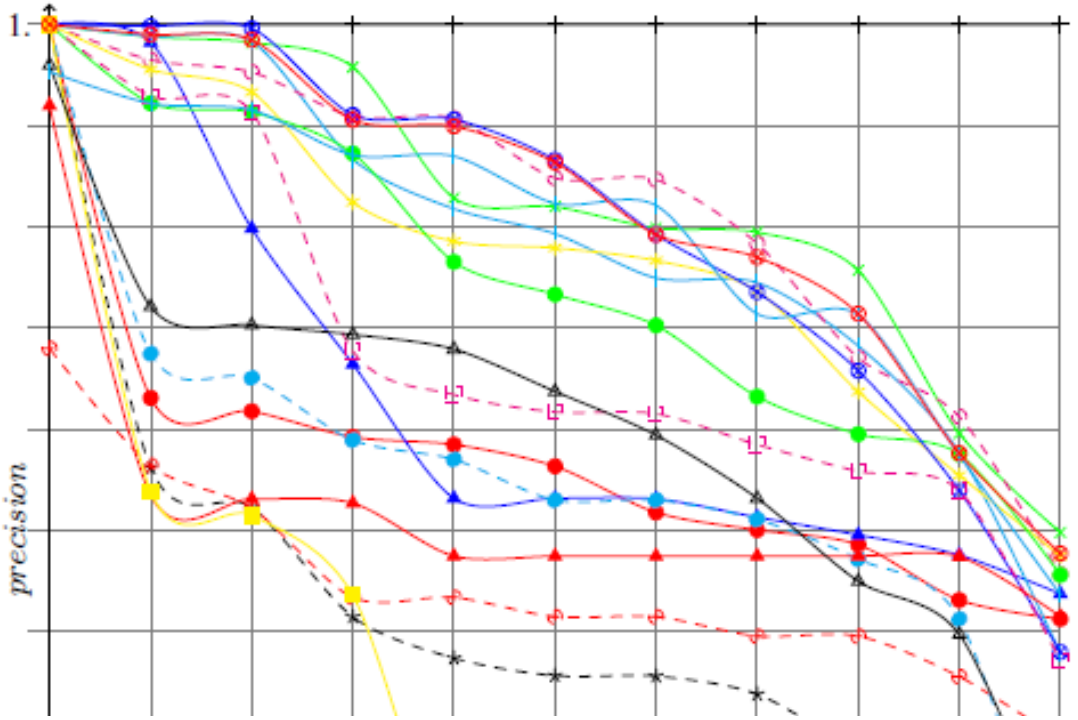
- Food

	RiMOM	Falcon-AO	Prior	COMA++	HMatch
Precision (taxonomical)	82%	83%⋆	68%	43%	48%
Precision (bio/chem)	85%⋆	80%	81%	76%	83%
Precision (miscellaneous)	78%	83%⋆	74%	70%	80%
Precision (all topics)	**81%**	**83%⋆**	**71%**	**54%**	**61%**

	RiMOM	Falcon-AO	Prior	COMA++	HMatch
Recall (all relations)	**50%**	**46%**	**45%**	**23%**	**46%**
Recall (only exactMatch)	71%	65%	64%	33%	65%

	RiMOM	Falcon-AO	Prior	COMA++	HMatch
F-measure (all rel. & top.)	**62%**	**59%**	**55%**	**33%**	**53%**

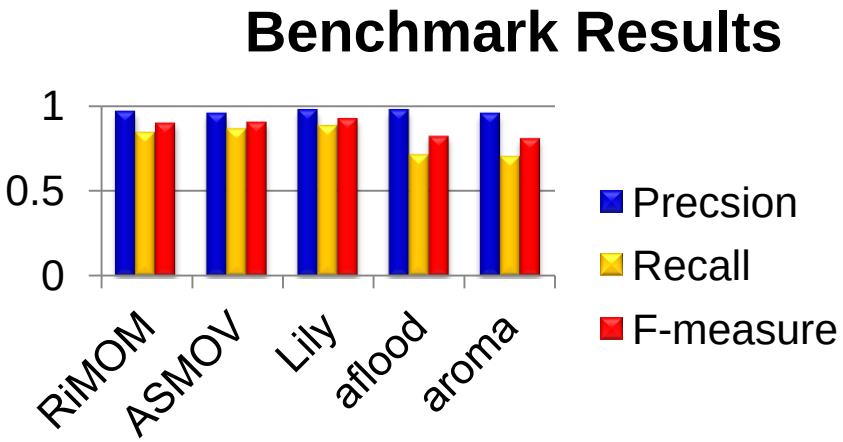
Results on OAEI 2007

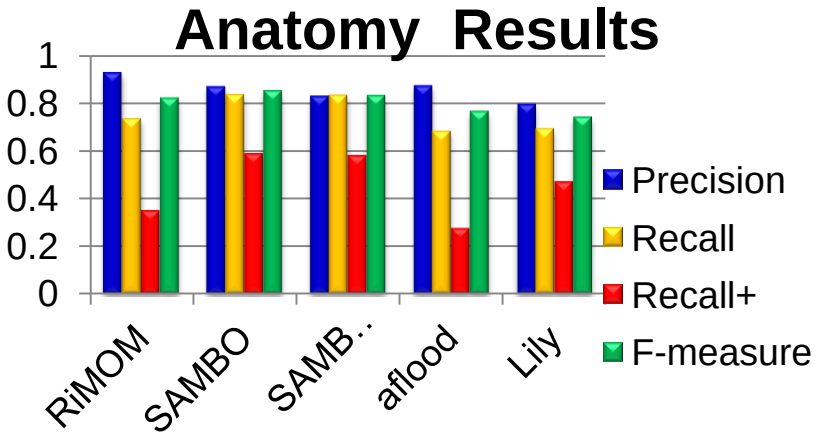


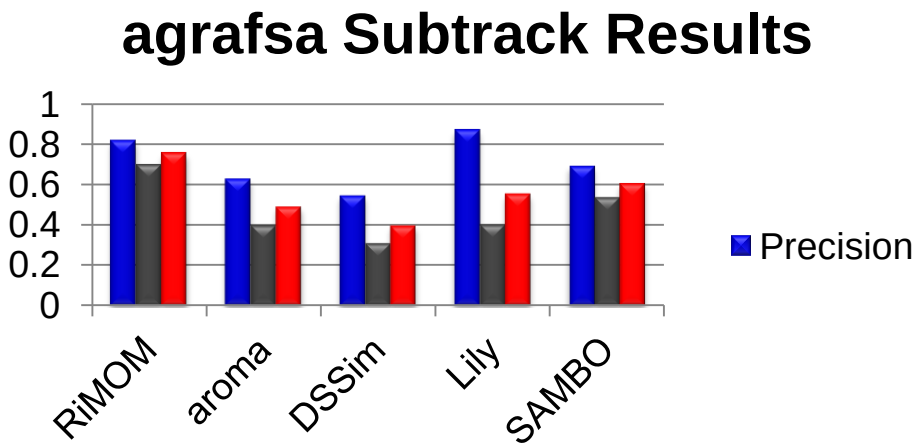
Year	2004				2005		2006		2007					
System	Fujitsu		PromptDiff		Falcon		RiMOM		ASMOV		Lily		RiMOM	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2xx	0.93	0.84	0.98	0.72	0.98	0.97	1.00	0.98	0.99	0.99	1.00	0.98	1.00	0.97
3xx	0.60	0.72	0.93	0.74	0.93	0.83	0.83	0.82	0.85	0.82	0.81	0.80	0.69	0.80
H-means	0.88	0.85	0.98	0.77	0.97	0.96	0.97	0.96	0.97	0.97	0.97	0.96	0.95	0.95

Table 4. Evolution of the best scores over the years on the basis of 2004 tests.

Result on OAEI 2008







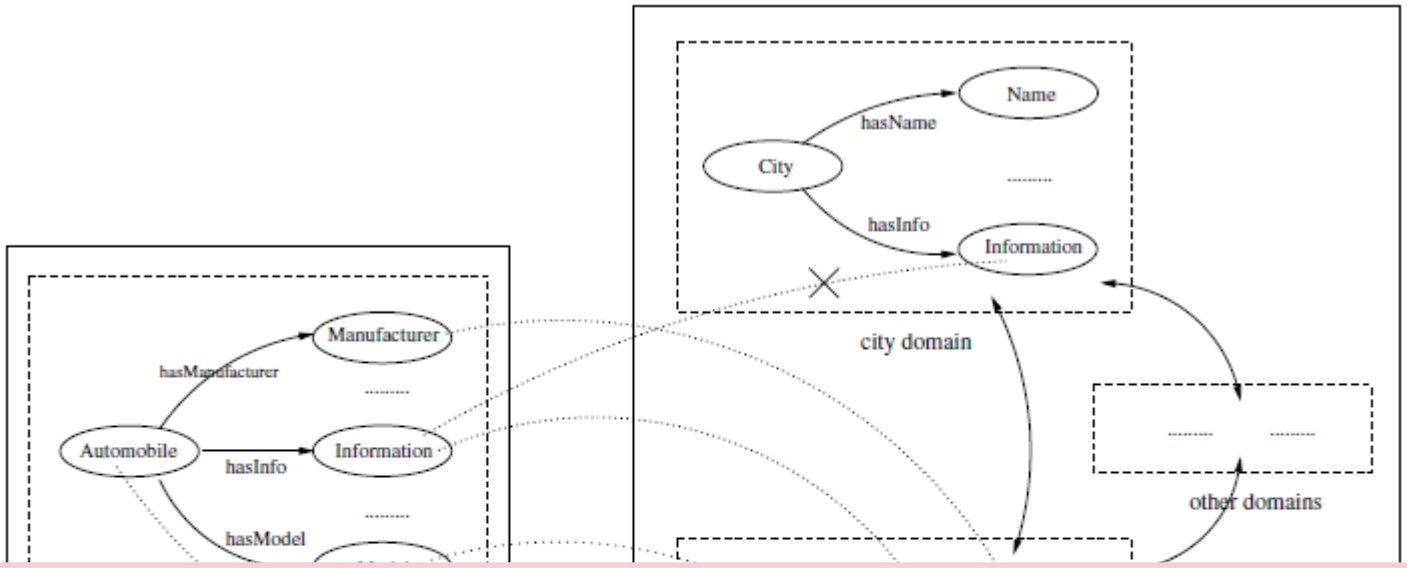
Experiences

- Structure information is very important in many alignment tasks for achieving high performance
- An effective method for combining the multiple strategies can enhance alignment performance
  - Investigate more factors to describe the characteristics of the ontologies
  - Exploit new strategies for ontology alignment

Outline

- Dynamic Multi-strategy Ontology Matching
- Unbalanced Ontology Matching
- Discussion

Unbalanced Ontology

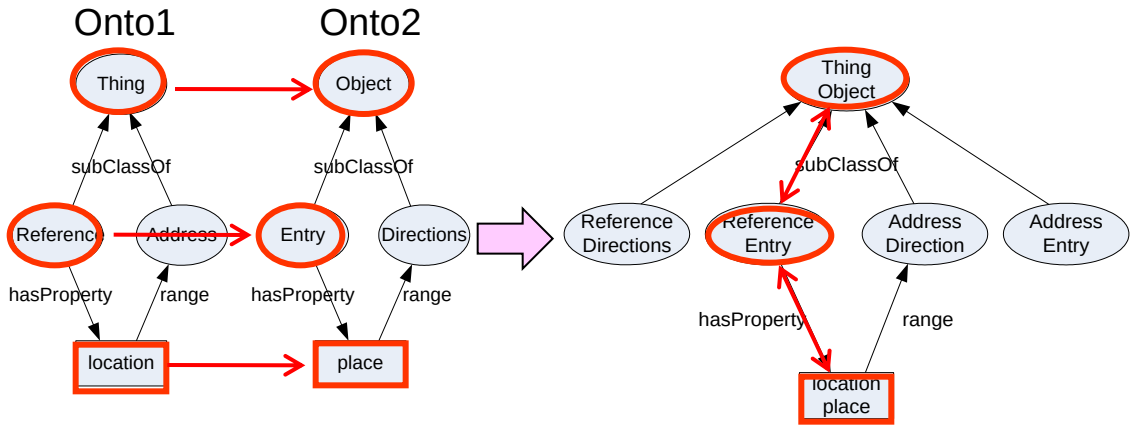


Several challenges:

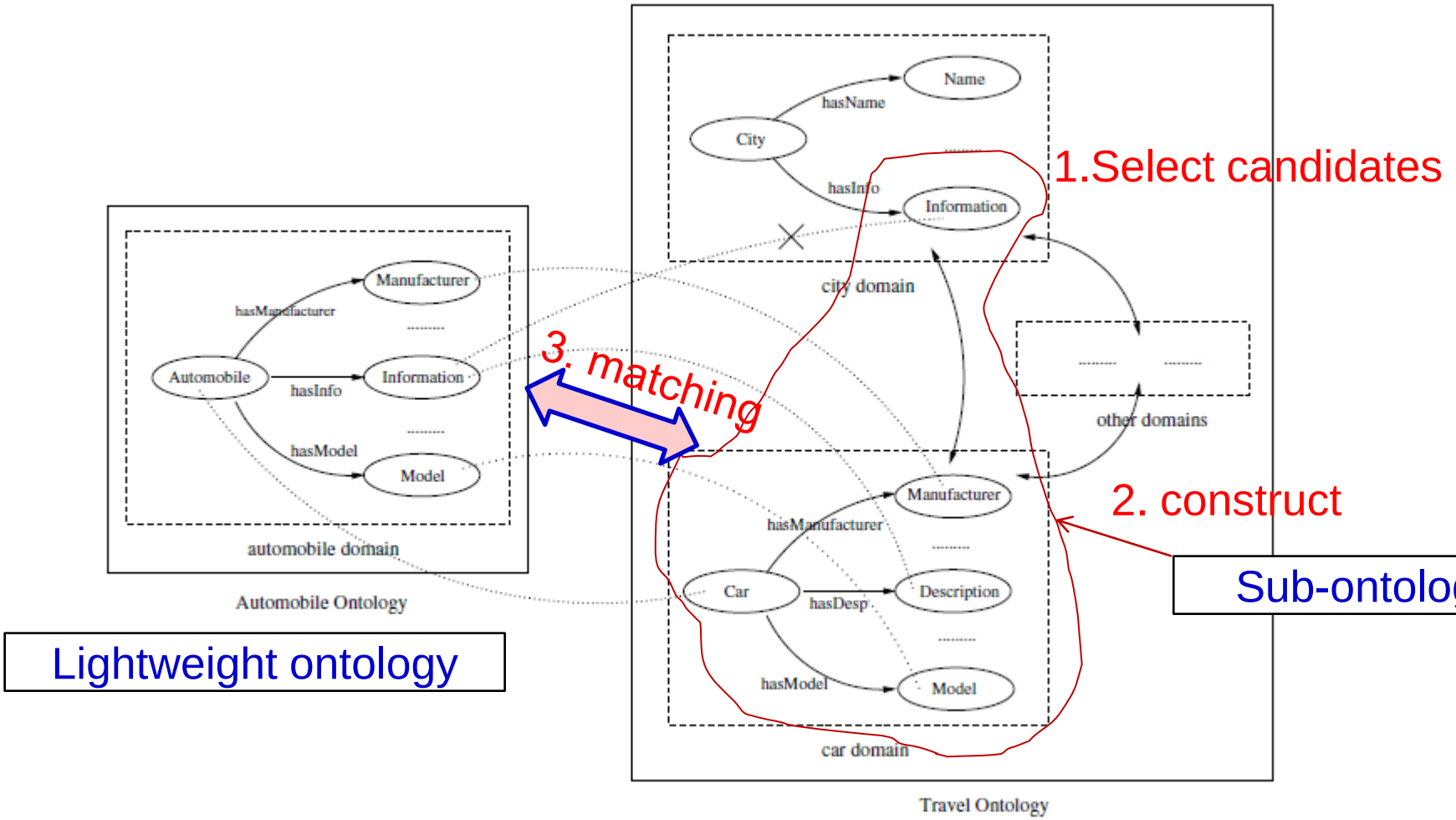
- **Single domain vs. multiple domains**
- **Small size vs. large-size ontology**

Key Problems

- Linguistic-based strategy
  - $|O_1| \times |O_2|$
- Structure-based strategy
  - In memory graphs
  - Iterative propagation



Our Approach



Step 1: Select Candidates

Algorithm 2 Selecting Candidate Concepts from O_h

1: **Input:** Lightweight ontology O_l, heavyweight ontology O_h.
2: **Output:** Concept set $C = \{c_i\}$, c_i i
3:
4: let C be empty
5: **for all** concept c_i in O_h **do**
6: **for all** concept c_j in O_l **do**
7: calculate similarity s_{ij} between
8: **if** $s_{ij} > \alpha$ **then**
9: $s_i = s_i + s_{ij}$
10: **end if**
11: **end for**
12: **if** $s_i > \beta$ **then**
13: add c_i to C
14: **end if**
15: **end for**

Complexity: $|O_1| \times |O_2|$

Edit-distance

$$s_{edit} = \frac{1}{1 + edit_distance(w_i, w_j)}$$

$$edit_distance(w_i, w_j) = \frac{|\{op_k\}|}{\max(length(w_i), length(w_j))}$$

e.g. site vs. cite

WordNet

$$s_{wordnet}(w_i, w_j) = \frac{2 \times \log p(s)}{\log p(s_i) + \log p(s_j)}$$

Similarity between c_i and O_l

Step 2: Construct Sub-ontology

Algorithm 3 Collecting Relevant Concepts

1: **Input:** Concept set C, ontology O_h, ontology O_l.
2: **Output:** Updated set C.
3:
4: **for all** c_i in C **do**
5: calculate relevance r_i of c_i ($|c_i - c_j| \leq 2$) .
6: **if** $r_i < \gamma$ **then**
7: remove c_i from C.
8: **end if**
9: **end for**

similarity

influence

$$r_i = s_i \times \varphi(c_i)$$

$$\varphi(c_i) = \Sigma \varphi_j(c_i)$$

$$\varphi_j(c_i) = s_j \times e^{-(|c_i - c_j|)^2} \quad (i \neq j)$$

Algorithm 4 Constructing Sub-ontology O_s

1: **Input:** Ontology O_h, ontology O_l, concept set C
2: **Output:** Sub-ontology O_s of O_h
3:
4: generate a set of connected sub-graphs $\{g_i\}$ from O_h.
5: calculate average vertex degree d_l of ontology O_l.
6: **for all** g_i **do**
7: calculate average degree d_i of g_i.
8: **if** $|d_l - d_i| > \tau$ **then**
9: remove g_i.
10: **end if**
11: **end for**
12: output remaining sub-graphs as ontology O_s.

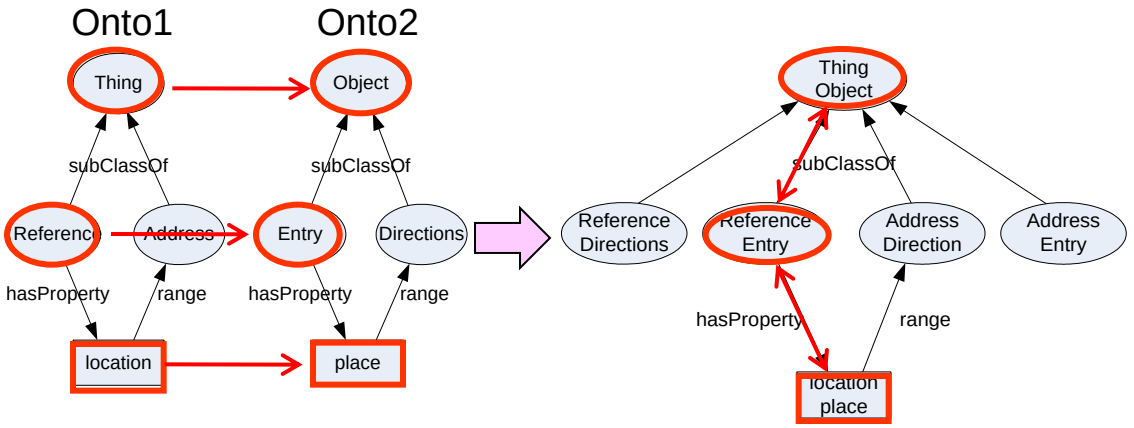
$$\frac{|E|}{|V|}$$

Step 3: Finding Matching Results

1. Given two ontologies O_1 and O_2, build a directed similarity graph G in which vertexes contain pairs of concepts in O_1 and O_2. In addition, if both concepts in one vertex have the same relation (in two ontologies respectively) with the concepts in anot edge is then constructed between the tv

2. Assign a weight w_{ij} to each edge $< v_i, v$ value of w_{ij} is set to be $1/n$ where n is of head vertex v_i.

3. Associate a similarity s_i^0 to each vertex be calculated by using other approaches, based methods.



4. Compute s_i^{n+1} for each vertex v_i with the following formula

$$s_i^{n+1} = s_i^n + \sum_j s_j^n \times w_{ij} + \sum_j s_j^n \times w_{ji}$$

5. Repeat step 4, until the difference between s_i^n and s_i^{n+1} is less than a given threshold.

Outline

- Dynamic Multi-strategy Ontology Matching
- Unbalanced Ontology Matching
  - Experimental Results
- Discussion

Data Set

- OAEI 2007
  - **GEMET**: (5,280) The European Environment Agency GEMET ontology.
  - **AGROVOC**: (28,439) AGROVOC thesaurus provided by Food and Agriculture Organization of the United Nations.
  - **NAL**: (42,326) The Agricultural thesaurus released by the National Agricultural Library.
- Evaluation Measures
  - Precision
  - Recall
  - F1-Measure
  - CPU Time

Data Statistics

Task	♯Matches	Lightweight Source	
		♯Concepts	Average Degree
ga_p_chem	14	46	4.67
ga_p_geo	23	43	5.51
ga_p_misc	28	89	4.28
ga_p_tax	21		
ga_p_nat	35		
ga_p_risk	21		
ga_r_agri	61		
ga_r_geo	87		

(a)GEM

Ontology	♯Concepts
GEMET	5280
AGROVOC	28439
NAL	42326

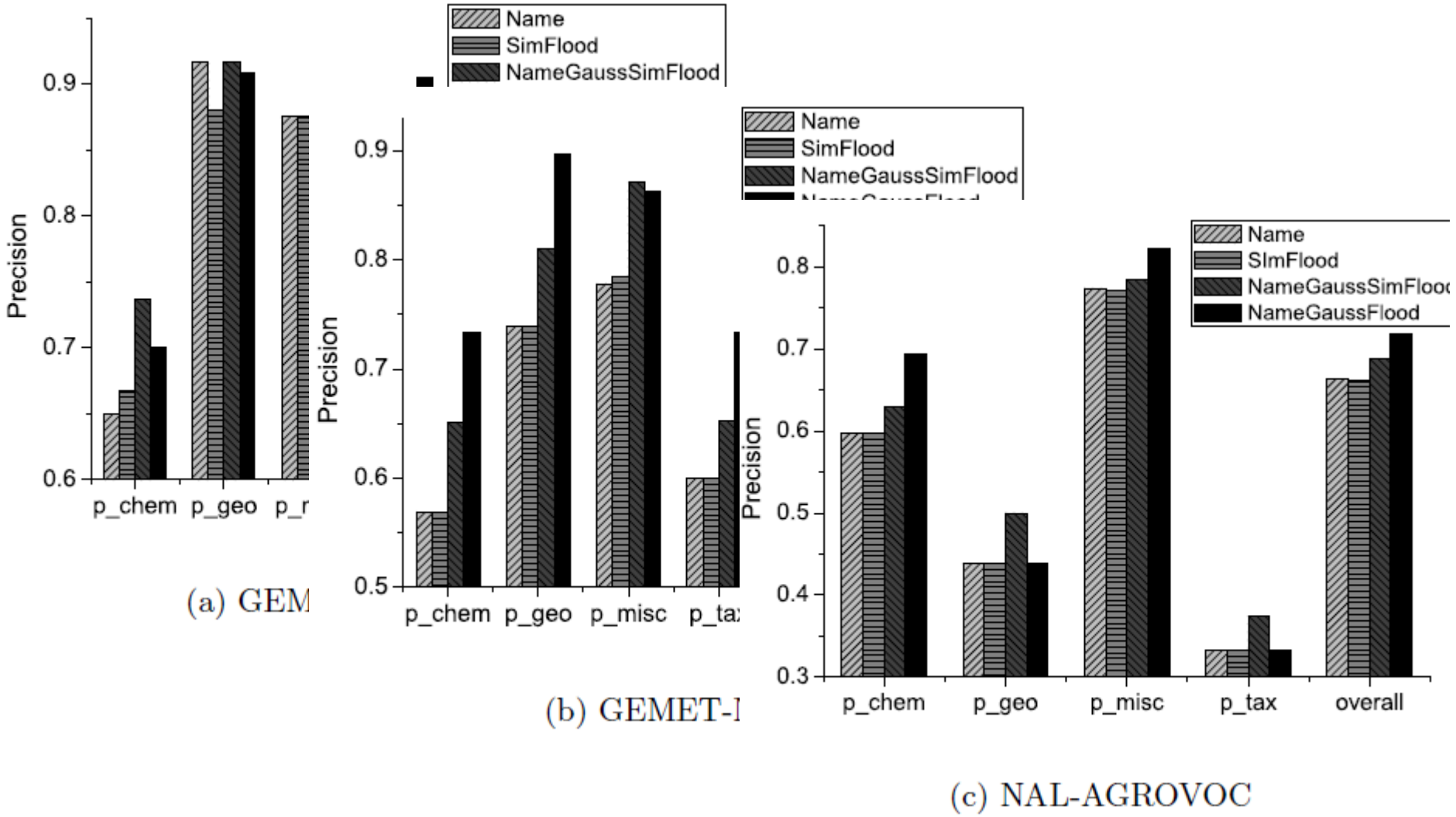
Task	♯Matches	Lightweight Source	
		♯Concepts	Average Degree
gn_p_chem	30	82	5.43
gn_p_geo	17	40	5
gn_p_misc	29		
gn_p_tax	15		
gn_p_nat	23		
gn_p_risk	30		
gn_r_agri	61		
gn_r_geo	77		

(b

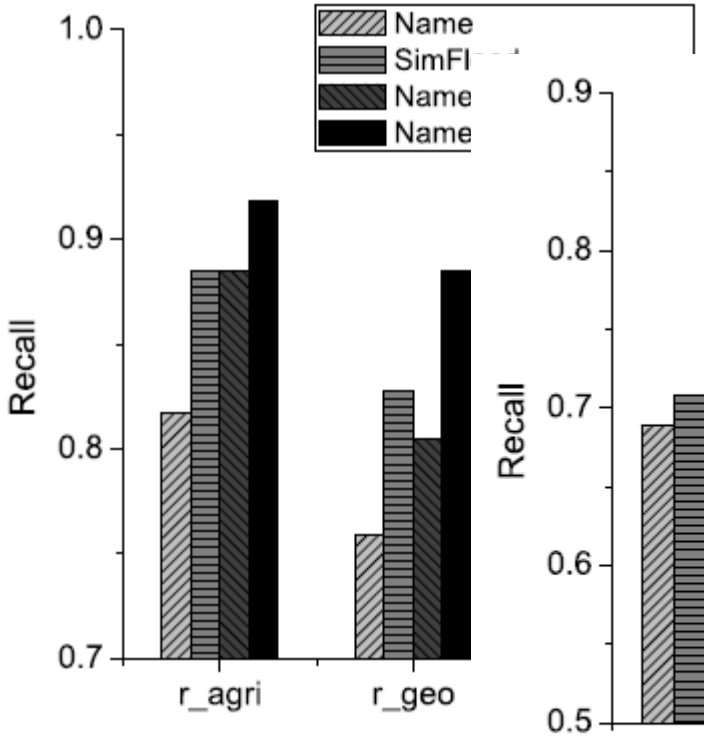
Task	♯Matches	Lightweight Source	
		♯Concepts	Average Degree
na_p_chem	141	283	1.92
na_p_geo	58	117	2.03
na_p_misc	231	575	1.80
na_p_tax	10	17	2
na_r_anim	10	39	2.36
na_r_rod	24	46	2.22
na_r_oaks	38	41	1.95
na_r_eur	62	71	3.44
na_r_geo	58	101	2.14

(c)NAL-AGROVOC

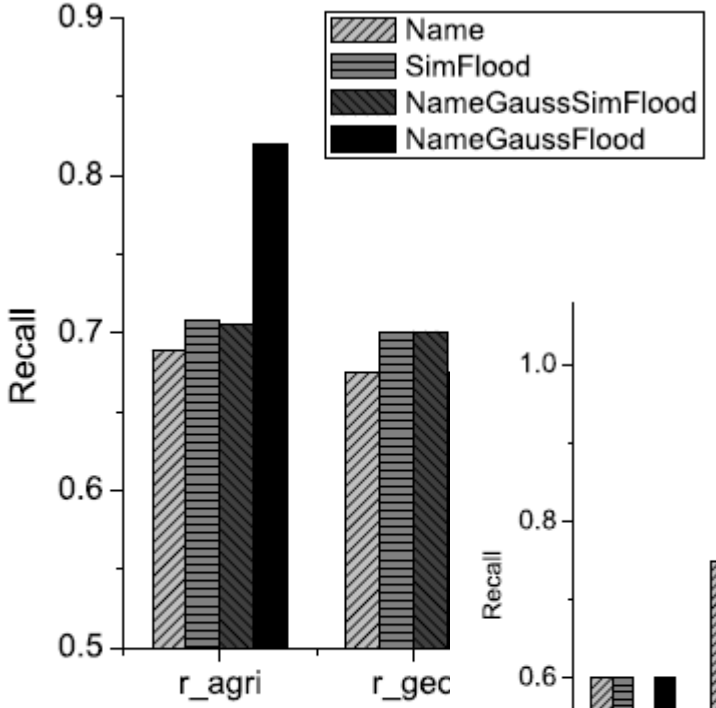
Precision


Name
SimFlood
NameGaussSimFlood
Precision
0.9
0.8
0.7
0.6
p_chem
p_geo
p_r
(a) GEM
Name
SimFlood
NameGaussSimFlood
Precision
0.9
0.8
0.7
0.6
0.5
p_chem
p_geo
p_misc
p_ta
(b) GEMET-I
Name
SImFlood
NameGaussSimFlood
NameGaussFlood
Precision
0.8
0.7
0.6
0.5
0.4
0.3
p_chem
p_geo
p_misc
p_tax
overall
(c) NAL-AGROVOC

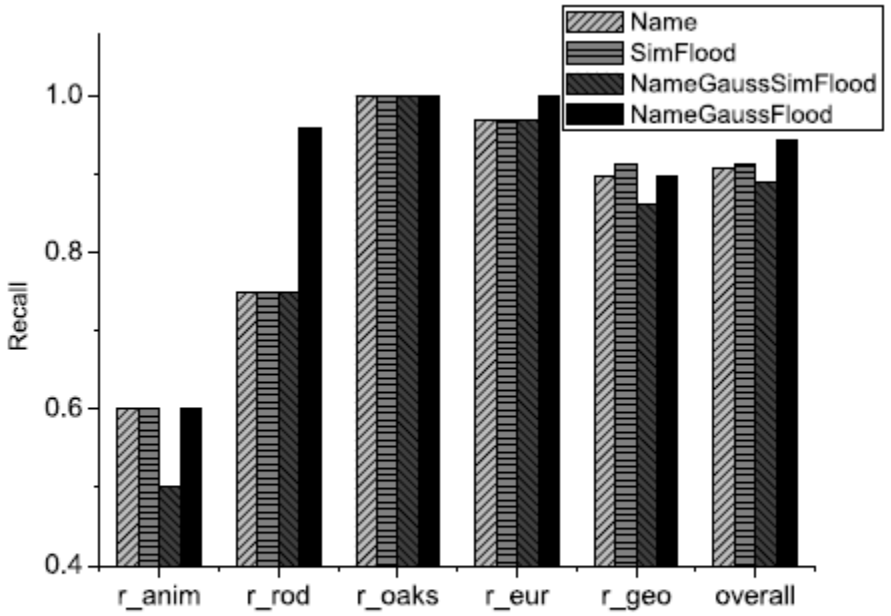
Recall



(a) GEMET-AGR(



(b) GEMET-N



(c) NAL-AGROVOC

F1-Measure

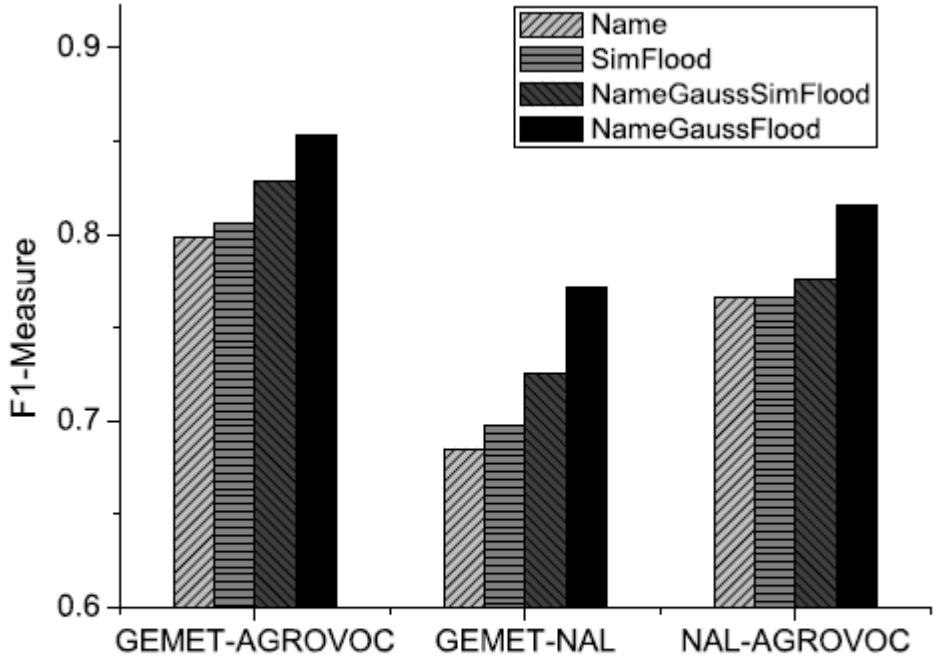
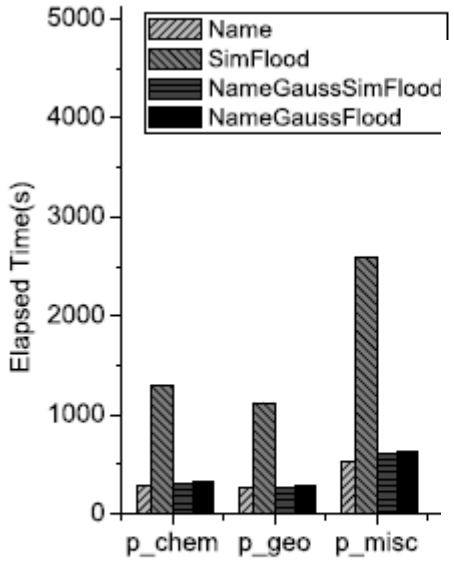
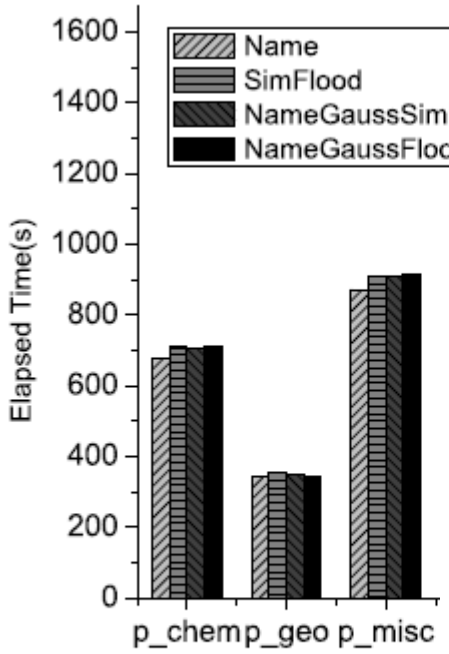


Figure 5: F1 Measure

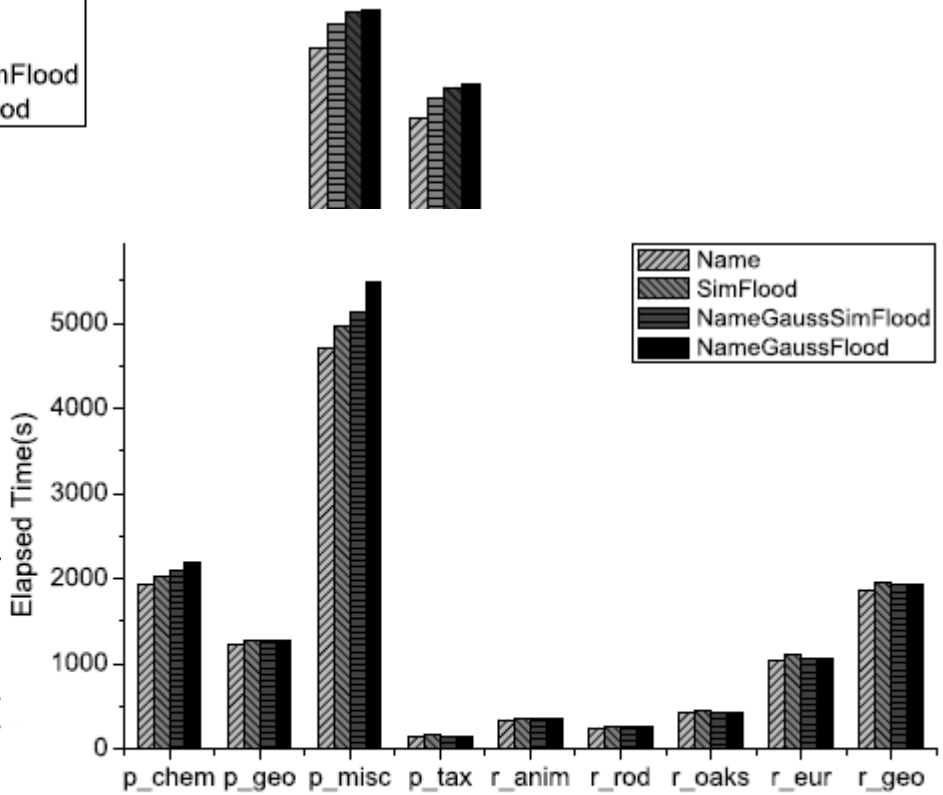
CPU Time



(a) GEME



(b) GE



(c) NAL-AGROVOC

Outline

- Dynamic Multi-strategy Ontology Matching
- Unbalanced Ontology Matching
- Discussion

Discussion

- Large-scale ontology matching
  - Both ontologies are very large
- Group ontology matching
  - A large number of sub ontologies
- Social ontology integration
  - Folksonomies
- Active learning for ontology matching
  - User interactions
- Beyond one-one alignment
- Beyond alignment

Related Publications

- Jie Tang, Juanzi Li, Bangyong Liang, Xiaotong Huang, Yi Li, and Kehong Wang. Using Bayesian Decision for Ontology Mapping. *Journal of Web Semantics*, Vol(4) 4:243-262, December 2006. **(Top 10 cited papers in JWS's history)**
- Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. RiMOM: A Dynamic Multi-Strategy Ontology Alignment Framework*. IEEE Transaction on Knowledge and Data Engineering (TKDE)*. August 2009 (vol. 21 no. 8) pp. 1218-1232. **(one of top cited papers among TKDE 2009's 100+ papers)**
- Qian Zhong, Hanyu Li, Juanzi Li, Guotong Xie, Jie Tang, and Lizhu Zhou. A Gauss Function based Approach for Unbalanced Ontology Matching. *In Proceedings of the 2009 ACM SIGMOD international conference on Management of data (SIGMOD'2009).* pp.669-680.
- Feng Shi, Juanzi Li, and Jie Tang. Actively Learning Ontology Matching via User Interaction. *In Proceedings of the 8th International Conference of Semantic Web (ISWC'2009).* pp. 585-600.

Thanks!

Q&A

HP:

http://keg.cs.tsinghua.edu.cn/persons/tj/