A session-based approach for aligning large ontologies

Patrick Lambrix, Rajaram Kaliyaperumal
Linköping University

Ontologies with overlapping information

- Use of multiple ontologies
  - custom-specific ontology + standard ontology
  - different views over same domain
  - overlapping domains

→ important to know the inter-ontology relationships

Ontology Alignment

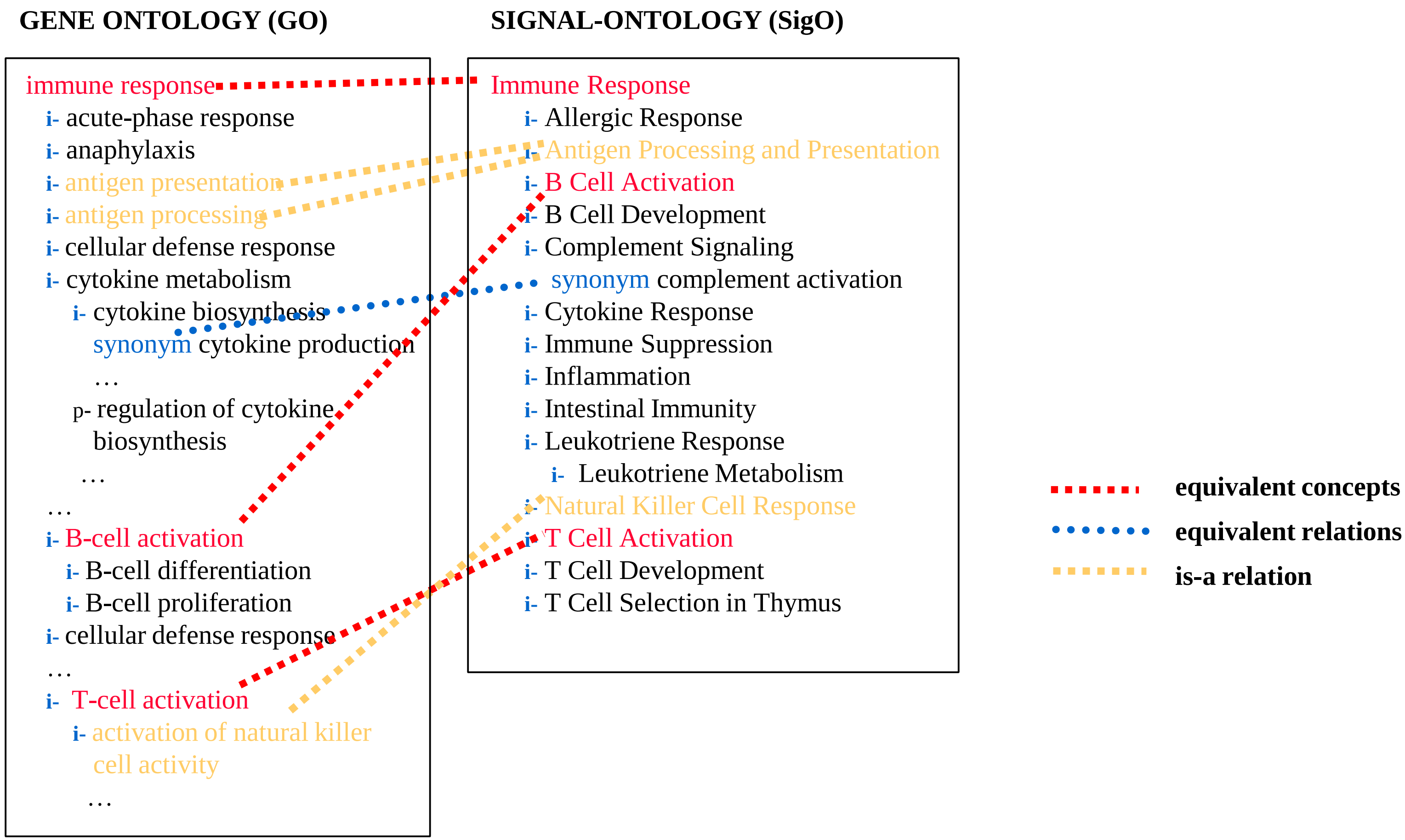
GENE ONTOLOGY (GO)

immune response
- i- acute-phase response
- i- anaphylaxis
- i- antigen presentation
- i- antigen processing
- i- cellular defense response
- i- cytokine metabolism
  - i- cytokine biosynthesis
    synonym cytokine production
    …
  - p- regulation of cytokine biosynthesis
  - …
- …
- i- B-cell activation
  - i- B-cell differentiation
  - i- B-cell proliferation
- i- cellular defense response
- …
- i- T-cell activation
  - i- activation of natural killer cell activity
  - …

SIGNAL-ONTOLOGY (SigO)

Immune Response
- i- Allergic Response
- i- Antigen Processing and Presentation
- i- B Cell Activation
- i- B Cell Development
- i- Complement Signaling
  synonym complement activation
- i- Cytokine Response
- i- Immune Suppression
- i- Inflammation
- i- Intestinal Immunity
- i- Leukotriene Response
  - i- Leukotriene Metabolism
- i- Natural Killer Cell Response
- i- T Cell Activation
- i- T Cell Development
- i- T Cell Selection in Thymus

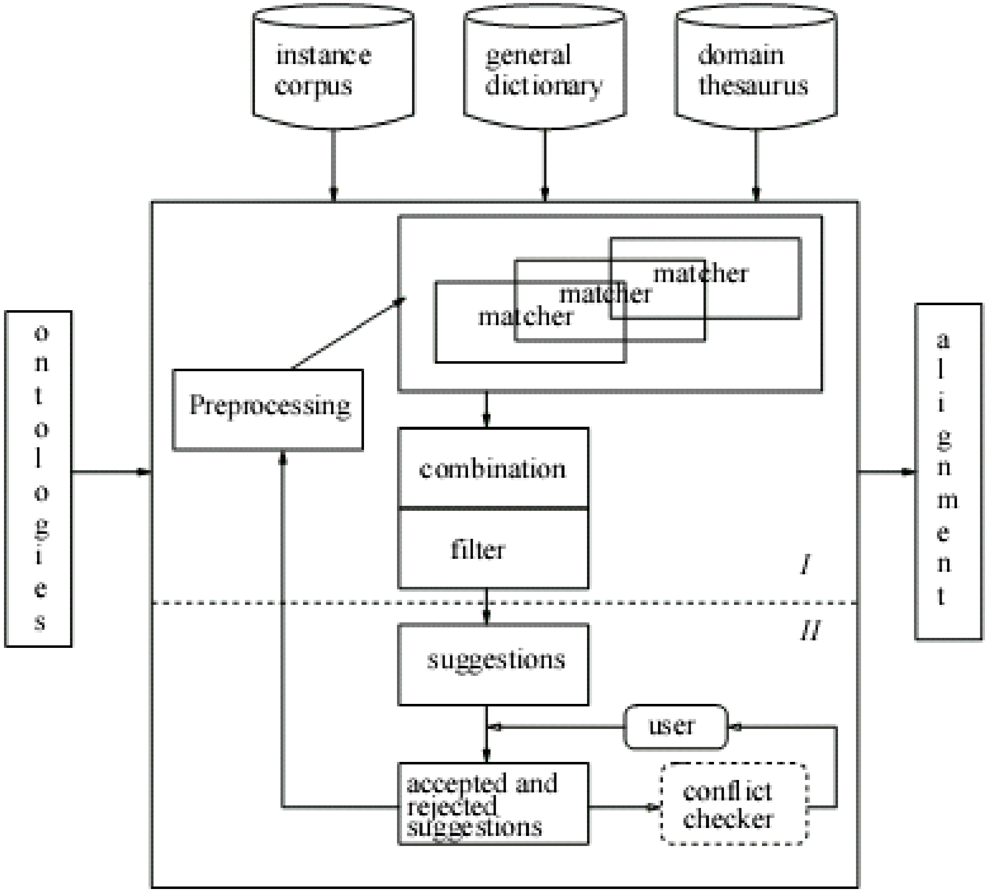
Ontology Alignment



define the relationships between the terms in different ontologies

Alignment framework

An Alignment Framework



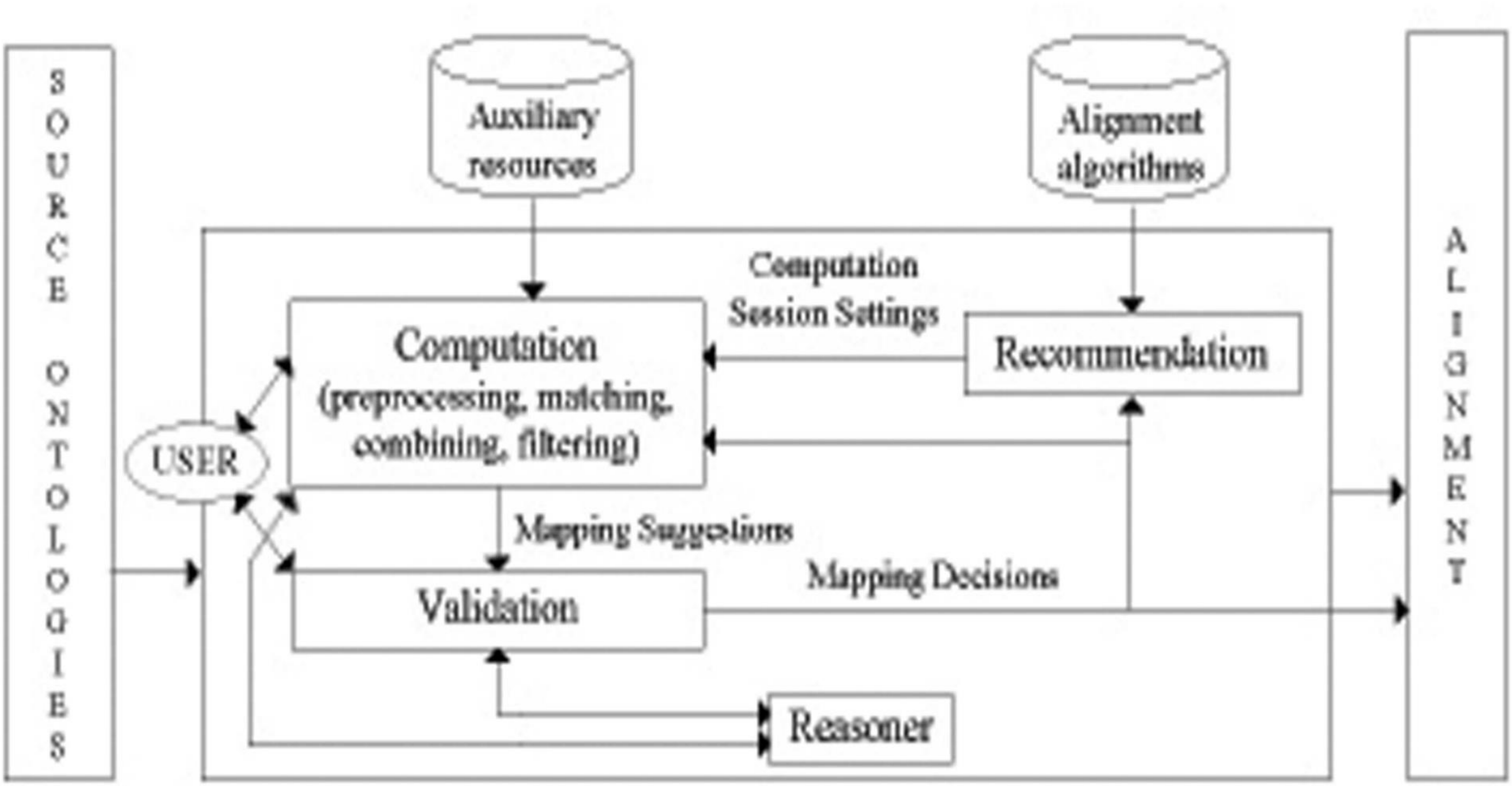
Challenges for aligning large ontologies

- Scalability
- Support for matcher selection, combination and tuning
- Use of background information
  - Partial results
- User involvement

(Shvaiko & Euzenat 2013)

Session-based framework

An Alignment Framework



Session-based approach

- Scalability – *interruptable sessions, partial computation, partial validation*
- Support for matcher selection, combination and tuning – *recommendation sessions*
- Use of background information –

  Use of partial results in computation and recommendation
- User involvement – *direct in setting process and validation, indirectly in computation and recommendation*

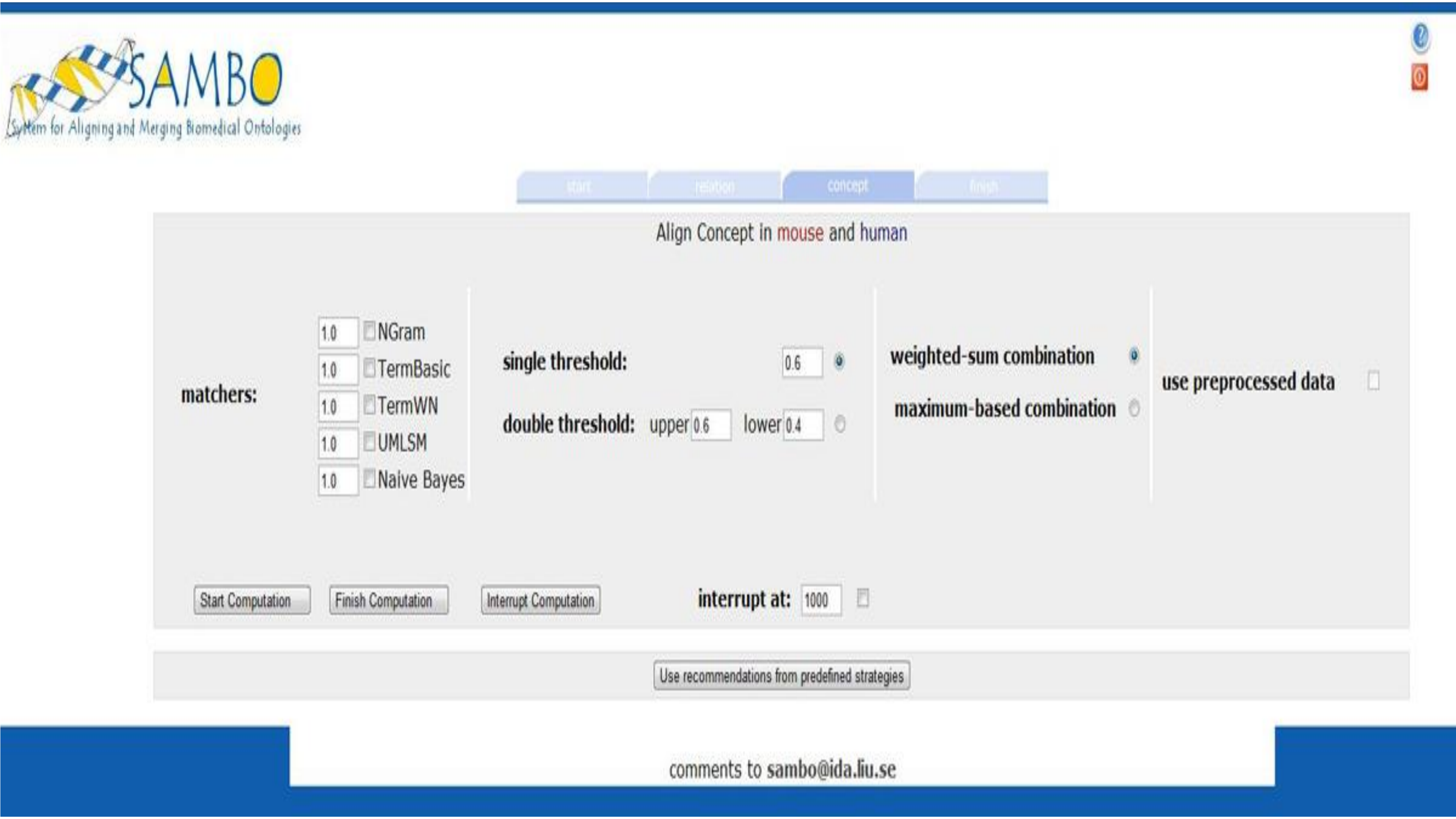
Implemented system

Databases

- Session management database
  - User, ontologies, validated mappings, non-validated mappings, …
  - Multiple sessions
- Similarity values database
  - Computation sessions, recommendation sessions
- Mapping decisions database
- Recommendation database

Implemented system – computation

Start of computation



Implemented system – computation
1. preprocessing

Use of PA in the preprocessing step

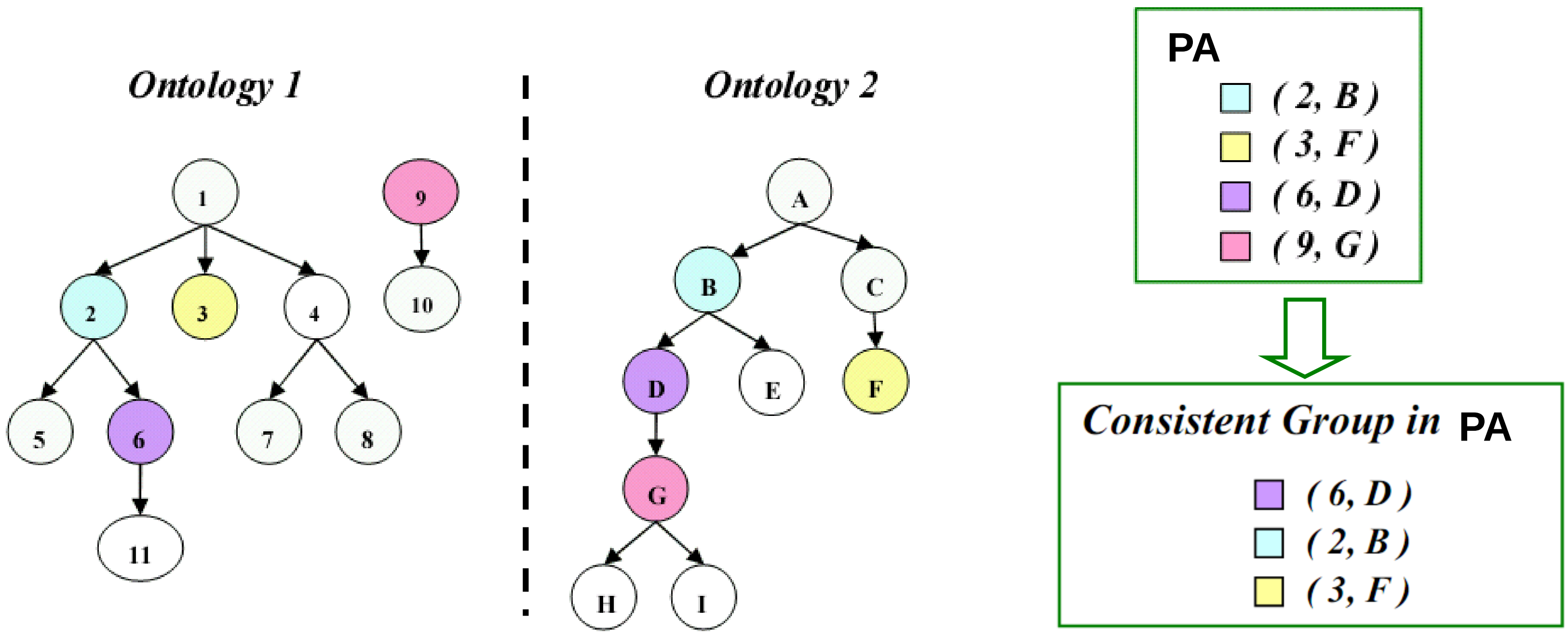
- Intuition

  During the preprocessing step, use mappings in PA to partition the ontologies into mappable groups.

(Lambrix & Liu 2009)

Use of PA in the preprocessing step

- Strategy
  - Find consistent group in PA
    - *if* (A,A') and (B,B') equivalence mappings in PA
      then A is-a B iff A' is-a B'
  - Partition ontologies into mappable groups before aligning

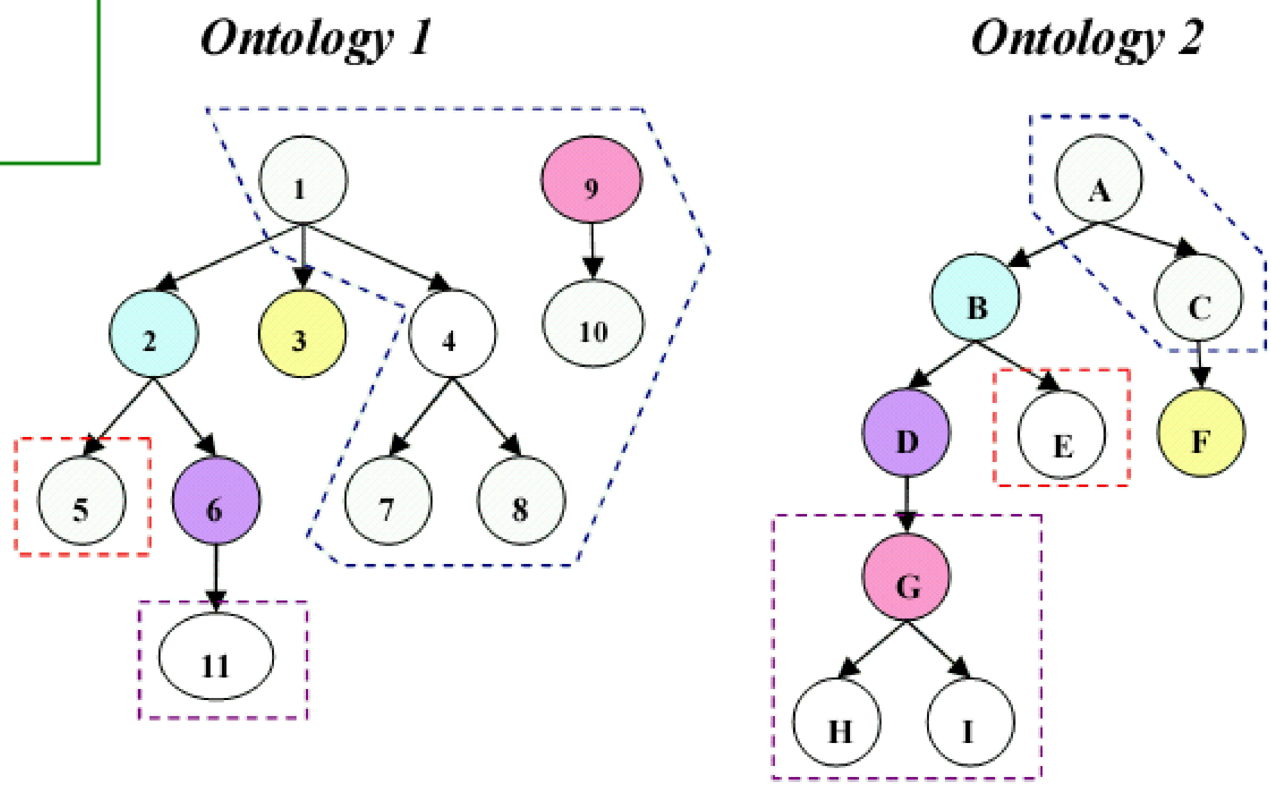


Use of PA in the preprocessing step

- Partition Results

Consistent Group in PA

- *(6, D)*
- *(2, B)*
- *(3, F)*



Implemented system – computation
2. matchers

Matchers

- N-gram (linguistic)
- TermBasic (linguistic)
- TermWN (linguistic + auxiliary)
- UMLS (auxiliary)
- Naive Bayes (instance-based)

(Lambrix & Tan 2006)

Implemented system – computation
3. combination strategies

Combination Strategies

- Weighted sum of similarity values of different matchers
- Maximum of similarity values of different matchers

Implemented system – computation
4. filtering strategies

Filtering Strategies

- Single threshold filtering
- Double threshold filtering

(Chen, Lambrix & Tan 2006)

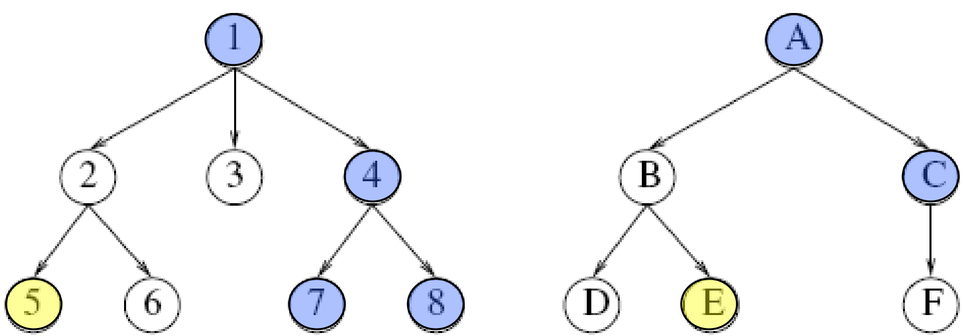
Filtering strategies

- Double threshold filtering

  (1) Pairs of concepts with similarity higher than or equal to **upper** threshold are mapping suggestions

  (2) Find consistent group among these mapping suggestions

  (3) Pairs of concepts with similarity between **lower** and **upper** thresholds are mapping suggestions if they make sense with respect to the structure of the ontologies and the suggestions according to consistent group



(2, B) *

(3, F)

(6, D) *

(4, C) *

upper-th - - - - - -

(5, C)

(5, E)

lower-th - - - - - -

……

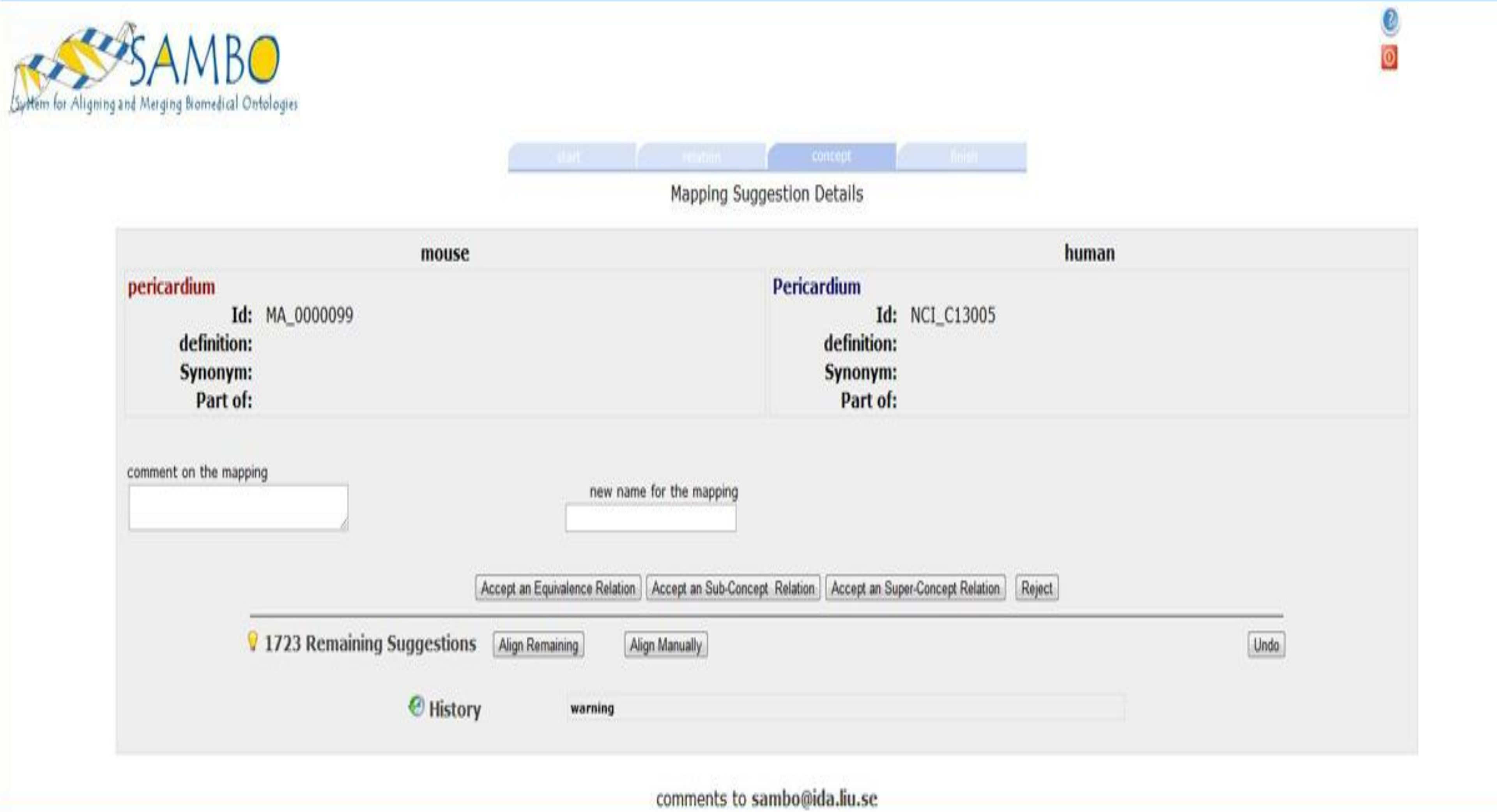
Filtering Strategies

- fPA – remove mappings suggestions conflicting with mappings in PA
- Double threshold filtering with PA
  - Use consistent group within PA

(Lambrix & Liu 2009)

Implemented system – validation

Validation


SAMBO
System for Aligning and Merging Biomedical Ontologies
start
relation
concept
finish
Mapping Suggestion Details
mouse
human
pericardium
Id: MA_0000099
definition:
Synonym:
Part of:
Pericardium
Id: NCI_C13005
definition:
Synonym:
Part of:
comment on the mapping
new name for the mapping
Accept an Equivalence Relation
Accept an Sub-Concept Relation
Accept an Super-Concept Relation
Reject
1723 Remaining Suggestions
Align Remaining
Align Manually
Undo
History
warning
comments to sambo@ida.liu.se

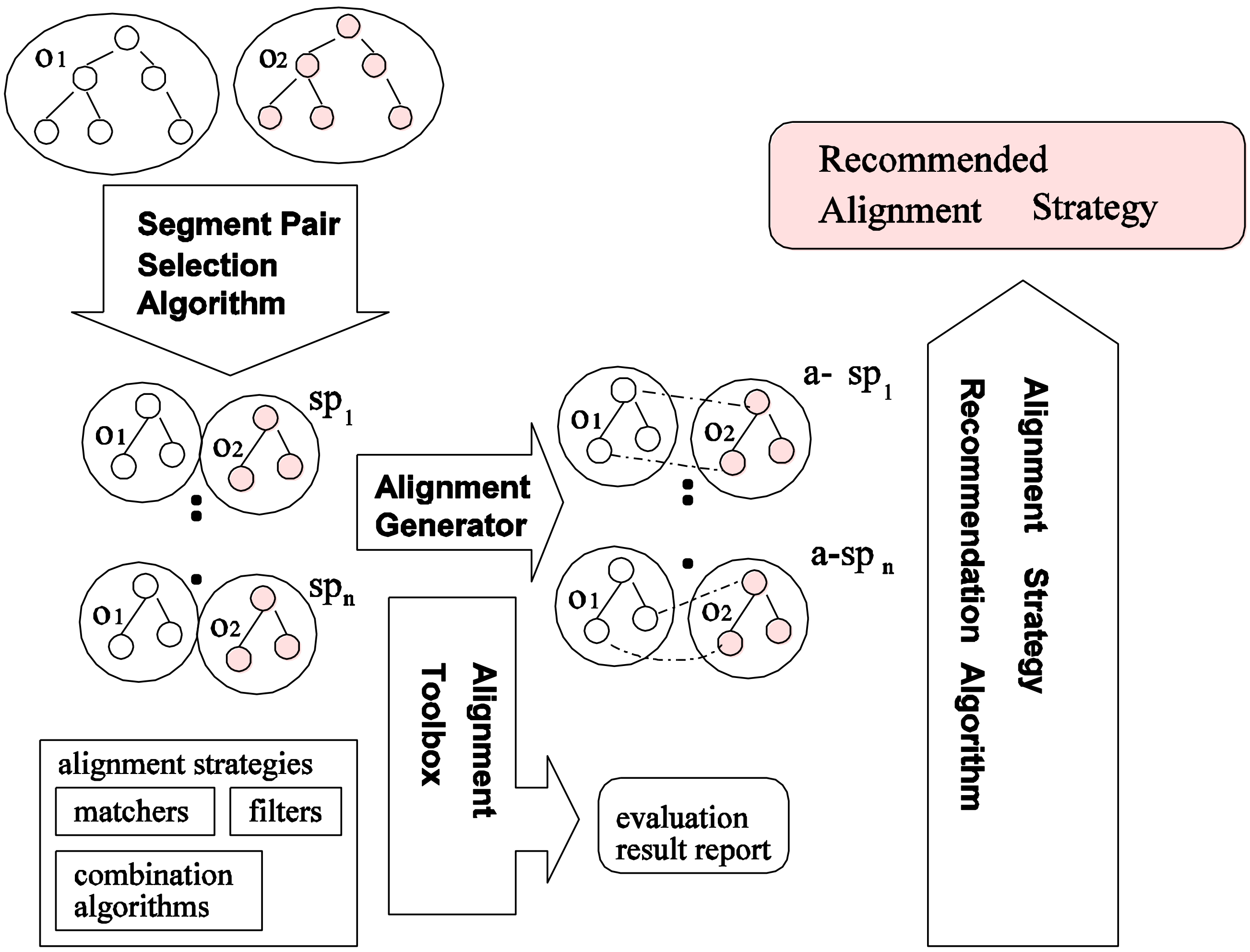
Implemented system – recommendation

Recommendation approach 1

- Select small segments of the ontologies
- Generate alignments for the segments (expert/oracle)
- Use and evaluate available alignment algorithms on the segments
- Recommend alignment algorithm based on evaluation on the segments

(Tan & Lambrix 2007)

Framework



o1
o2
Segment Pair Selection Algorithm
sp1
spn
Alignment Generator
a- sp1
a-sp n
Recommended Alignment Strategy
Alignment Strategy Recommendation Algorithm
Alignment Toolbox
alignment strategies
matchers
filters
combination algorithms
evaluation result report

Recommendation approach 2

- Evaluate available alignment algorithms on previous validation decisions
- Recommend alignment algorithm based on evaluation on the validation decisions

Recommendation approach 3

- Select small segments of the ontologies
- Evaluate available alignment algorithms on the segments based on previous validation decisions
- Recommend alignment algorithm based on evaluation on the segments

Recommendation approaches

- Approach 1
  - based on full knowledge of mappings in validated segments
  - Need domain expert/oracle
  - Good performance for segments does not necessarily lead to good performance for ontologies

- Approaches 2 and 3
  - No full knowledge of mappings may be available for any parts of the ontologies
  - No need for domain expert/oracle during recommendation
  - Validation decisions can come from different parts of the ontologies

Experiments

Experiments

- As an ontology alignment system
- For evaluation of ontology alignment strategies

Experiments

- OAEI 2011 Anatomy track
  - AMA, 2737 concepts
  - NCI-A, 3298 concepts
  - Reference alignment, 1516 equivalence mappings
- 5 matchers, 2 combination,
  2 filter / 6 thresholds → 4872 strategies

Top 10 strategies

matchers	weights	threshold	correct suggestions	wrong suggestions	F^c	Sim2
TermBasic;UMLSM	1;1	0.4;0.7	1223	101	0.8612	0.7563
TermWN;UMLSM;NaiveBayes;n-gram	1;2;2;1	0.3;0.5	1223	101	0.8612	0.7563
n-gram;TermBasic;UMLSM	1;1;2	0.5;0.8	1192	63	0.8603	0.7549
n-gram;UMLSM	1;1	0.5;0.8	1195	67	0.8603	0.7548
UMLSM;NaiveBayes;TermWN	2;1;2	0.4;0.6	1203	78	0.8602	0.7547
UMLSM;NaiveBayes;n-gram;TermBasic	2;1;1;1	0.4;0.6	1199	73	0.8601	0.7545
n-gram;TermBasic;UMLSM	1;2;2	0.5;0.8	1181	50	0.8598	0.7541
UMLSM;NaiveBayes;TermBasic	2;1;2	0.4;0.6	1194	68	0.8596	0.7537
UMLSM;NaiveBayes;n-gram;TermBasic	2;2;1;1	0.3;0.5	1221	104	0.8595	0.7537
UMLSM;NaiveBayes;TermBasic	2;1;1	0.5;0.6	1187	60	0.8592	0.7531

Test strategies

strategy	matchers	weights	threshold	suggestions	F^c	Sim2
AS1	*TermBasic;UMLSM*	1;1	0.4;0.7	1324	0.86	0.75
AS2	*TermWN;n-gram;NaiveBayes*	2;1;1	0.5	1824	0.65	0.48
AS3	*n-gram;TermBasic;UMLSM*	1;1;2	0.3	4061	0.48	0.32

Matcher computation time

	n-gram		*NaiveBayes*	
number of pairs	without previous values stored	with previous values stored	without previous values stored	with previous values stored
902,662	2.59		196.15	
1,805,324	5.08	3.98	149.95	84.05
4,513,310	12.73	10.78	418.49	265.87
6,769,965	19.19	13.83	645.71	212.35
9,026,626	25.85	17.32	790.74	207.64

- performance gains up to 25%

Filter using validated correct mappings

processed	AS1	AS2	AS3
500	20	107	156
1000	26	58	288
1300	4	20	20

- Removal of mapping suggestions conflicting with validated correct mappings

  → reduce unnecessary user interaction

Double threshold filter using validated correct mappings

processed	AS1 suggestions removed	AS2 suggestions removed	AS3 suggestions removed	AS1 correct removed	AS2 correct removed	AS3 correct removed
500	0/2	134/113	244/279	0/0	12/1	9/1
1000	1/0	52/47	532/470	1/0	1/0	22/4
1300	0/2	43/35	443/276	0/0	9/2	21/3

- Removal of suggestions using double threshold filtering with validated correct mappings
- Original ontologies / missing is-a relations added

Recommendations

- Session-independent, segment pairs, oracle
  - No change during process
  - Dependent on original segments

Recommendations

- Session-dependent, validation decisions
  - Not good for AS1, double threshold filtering
  - AS1 suggested for AS3
- Session-dependent, segments, validation decisions
  - Not good for AS1, lack of wrong suggestions
  - Recommendation improves with more validations

Conclusion

- Session-based framework
  - Computation, validation, recommendation
  - Addressed several challenges
- System
- Experiments

Future work

- Use of validation results in computation and recommendation
- Recommendation strategies