UNIVERSITÀ DEGLI STUDI
DI TRENTO

# S-Match:
# an Algorithm and an Implementation of Semantic Matching

**Fausto Giunchiglia**

**work in collaboration with**
**Pavel Shvaiko and Mikalai Yatskevich**

**July 2004, Hannover, Germany**



# Outline

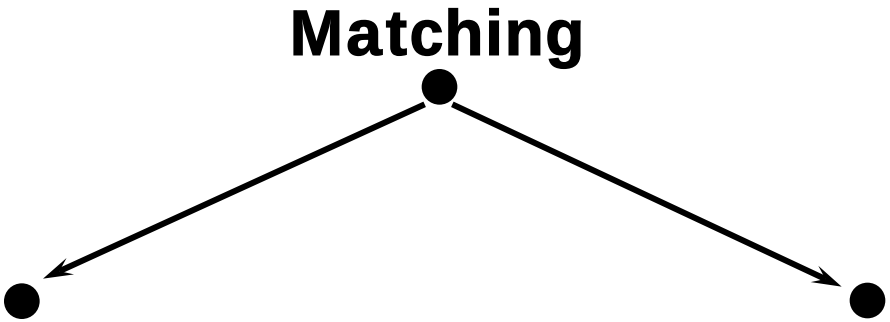
- Semantic Matching
- The S-Match Algorithm
- Element Level Semantic Matching
- The S-Match System
- A Comparative Evaluation
- Future Work



# Semantic Matching

# Matching

**Matching:** given two graph-like structures (e.g., concept hierarchies or ontologies), produce a mapping between the nodes of the graphs that semantically correspond to each other




## Syntactic Matching

- **Relations** are computed between **labels** at nodes
- **R = {x∈[0,1]}**

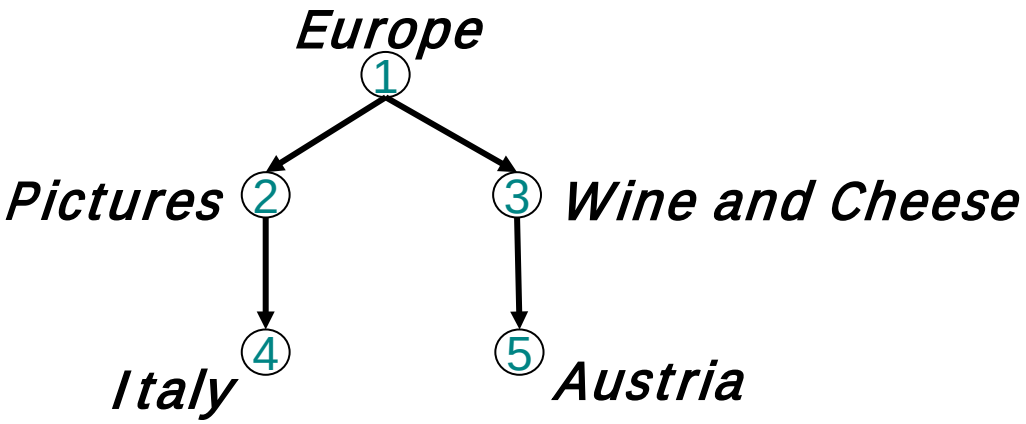
**Note:** all previous systems are syntactic…

## Semantic Matching

- **Relations are** computed between **concepts** at nodes
- **R = { =, ⊒ , ⊑ , ⊥, ⊓}**

**Note:** First implementation CTXmatch [Bouquet et *al. 2003*]

# Concept of a Label



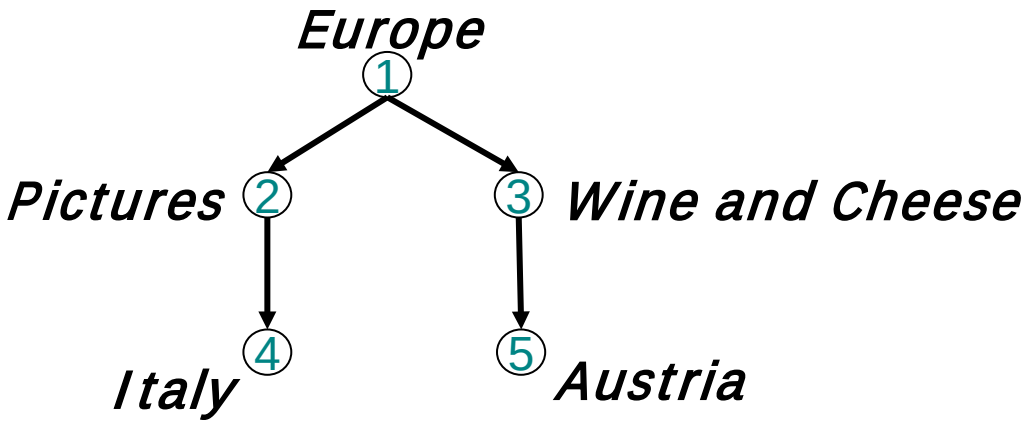

- **The idea:**
  - **Labels in classification hierarchies are used to define the set of documents one would like to classify under the node holding the label**
  - **A label has an intended meaning, which is what this label means in the world**

**Concept of a label is the set of documents that are about what the label means in the world**

# Concept of a Node

- **Observations:**
  - the semantics of a label are the real world semantics
  - the semantics of the concept of a label are in the space of documents
  - the relation being that the documents in the extension of the concept of a label are about what the label means in the real world

- **The idea:** Trees add structure which allows us to perform the classification of documents more effectively




**Concept of a node** is the set of documents that we would classify under this node, given it has a certain label and it is positioned in a certain place in the tree



# Semantic Matching

**Mapping element is a 4-tuple** $< ID_{ij}, n1_i, n2_j, R >$**, where**

- $ID_{ij}$ **is a unique identifier of the given mapping element;**
- $n1_i$ **is the *i-th* node of the first graph;**
- $n2_j$ **is the *j-th* node of the second graph;**
- *R* **specifies a semantic relation between the concepts at the given nodes**

**Computed *R's*, listed in the decreasing binding strength order:**

equivalence { = };

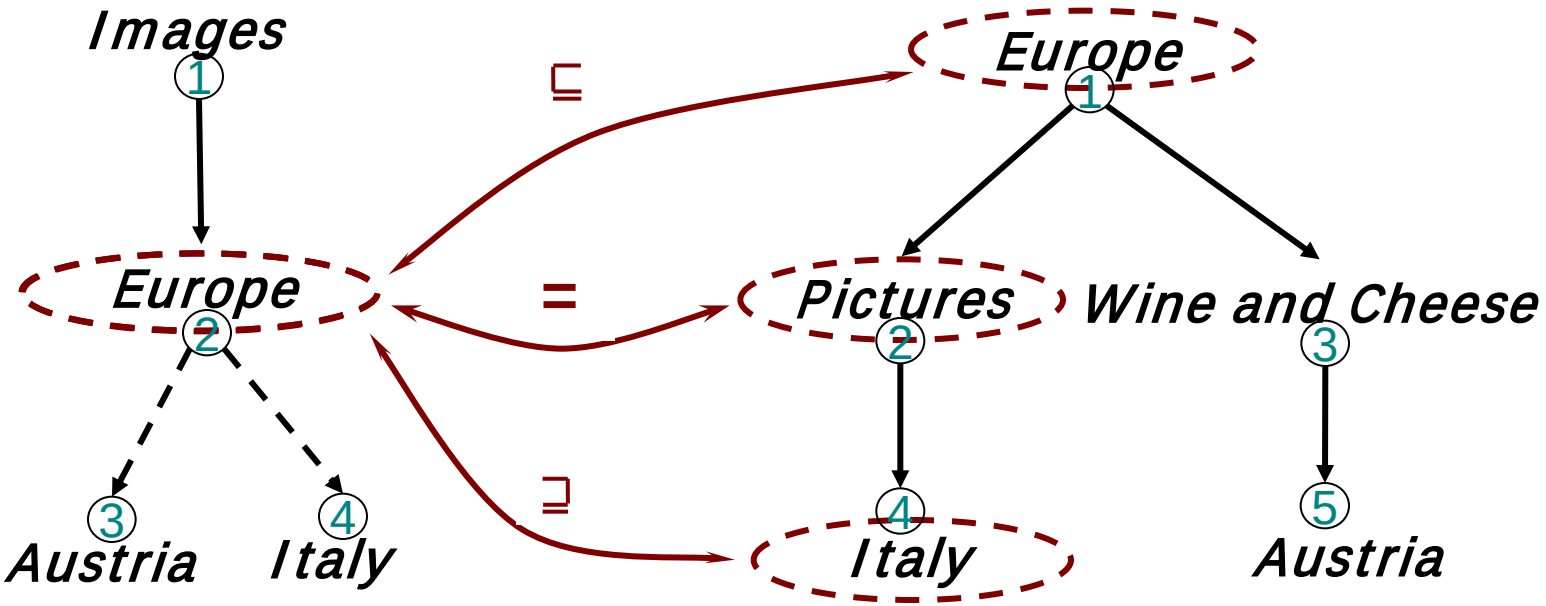
more general/specific { $\sqsupseteq$ , $\sqsubseteq$ };

mismatch { $\perp$ };

overlapping { $\sqcap$ }.

**Semantic Matching: Given two graphs *G1* and *G2*, for any node $n1_i \in G1$, find the strongest semantic relation *R'* holding with node $n2_j \in G2$**

# Example




$< ID_{22}, 2, 2, = >$

$< ID_{21}, 2, 1, \sqsubseteq >$

$< ID_{24}, 2, 4, \sqsupseteq >$

$\Rightarrow < ID_{22}, 2, 2, = >$

Algo
Step 4



# The S-Match Algorithm



# Four Macro Steps

**Given two labeled trees T1 and T2, do:**

1. For all labels in T1 and T2 compute ***concepts at labels***
2. For all nodes in T1 and T2 compute ***concepts at nodes***
3. For all pairs of labels in T1 and T2 compute **relations between concepts at labels**
4. For all pairs of nodes in T1 and T2 compute **relations between concepts at nodes**

Steps 1 and 2 constitute the preprocessing phase, and are executed once and each time after the schema/ontology is changed **(OFF- LINE part)**

Steps 3 and 4 constitute the matching phase, and are executed every time the two schemas/ontologies are to be matched **(ON - LINE part)**

# Step 1: compute concepts at labels

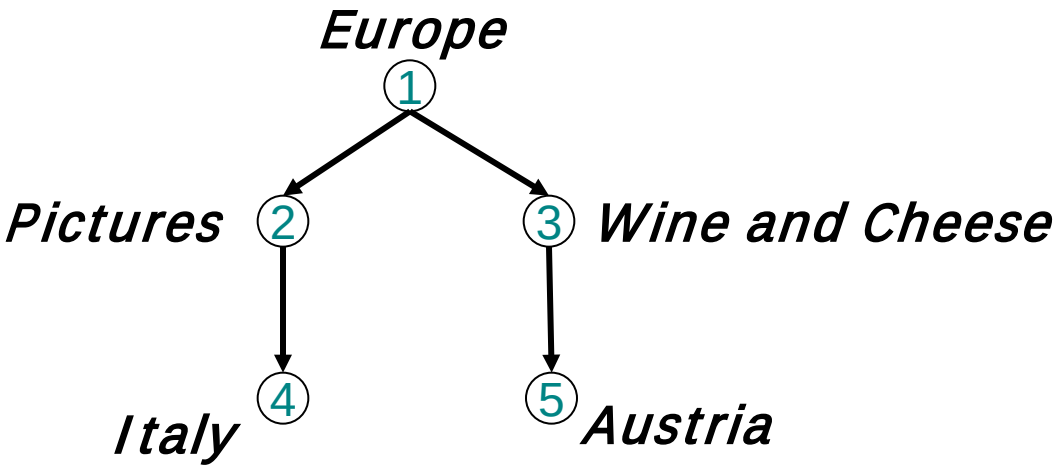
- **The idea:**
  - Translate natural language expressions into internal formal language
  - Compute concepts based on possible *senses* of words in a label and their interrelations

- **Preprocessing:**
  - **Tokenization.** Labels (according to punctuation, spaces, etc.) are parsed into tokens. E.g., Wine and Cheese $\rightarrow$ <Wine, and, Cheese>;
  - **Lemmatization.** Tokens are morphologically analyzed in order to find all their possible basic forms. E.g., Images $\rightarrow$ Image;
  - **Building atomic concepts.** An oracle (WordNet) is used to extract senses of lemmatized tokens. E.g., Image has 8 senses, 7 as a noun and 1 as a verb;
  - **Building complex concepts.** Prepositions, conjunctions, etc. are translated into logical connectives and used to build complex concepts out of the atomic concepts

    E.g., $C_{Wine\ and\ Cheese}$ = <Wine, U($WN_{Wine}$)> $\sqcup$ <Cheese, U($WN_{Cheese}$)>,
    where U is a union of the senses that WordNet attaches to lemmatized tokens

# Step 2: compute concepts at nodes

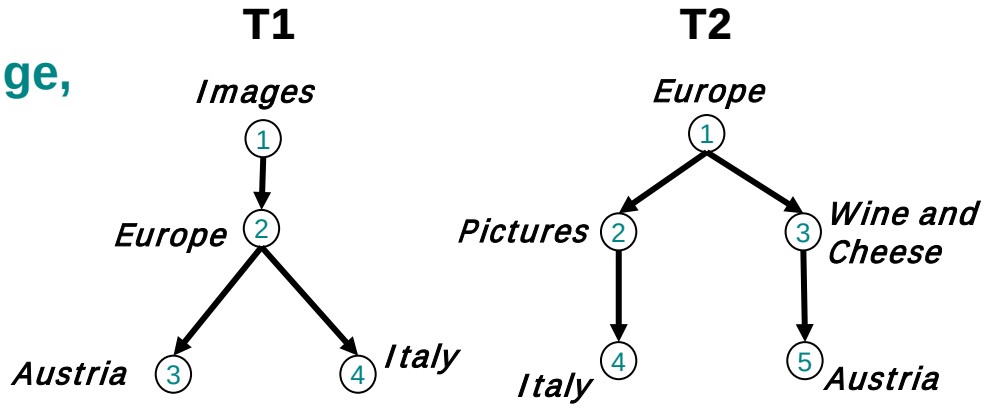
- **The idea:** extend **concepts at labels** by capturing the knowledge residing in a structure of a graph in order to define a context in which the given **concept at a label** occurs
- **Computation:** **Concept at a node** for some node *n* is computed as an intersection of **concepts at labels** located above the given node, including the node itself




$$C_4 = C_{Europe} \sqcap C_{Pictures} \sqcap C_{Italy}$$

# Step 3: compute relations between concepts at labels

- **The idea:** Exploit a priori knowledge, e.g., lexical, domain knowledge with the help of element level semantic matchers




- **Results of step 3:**

| T1 \ T2 | $C_{Europe}$ | $C_{Pictures}$ | $C_{Wine}$ | $C_{Cheese}$ | $C_{Italy}$ | $C_{Austria}$ |
|---|---|---|---|---|---|---|
| $C_{Images}$ | | = | | | | |
| $C_{Europe}$ | = | | | | $\sqsupseteq$ | $\sqsupseteq$ |
| $C_{Austria}$ | $\sqsubseteq$ | | | | $\bot$ | = |
| $C_{Italy}$ | $\sqsubseteq$ | | | | = | $\bot$ |



# Step 4: compute relations between concepts at nodes

**The idea:** Reduce the matching problem to a validity problem

- We take the relations between **concepts at labels** computed in step 3 as axioms (**Context**) for reasoning about **rel**ations between **concepts at nodes**. ***rel*** **= { =, ⊒ , ⊑ , ⊥, ⊓ }.**
- Construct the propositional formula ($C1_i$ in Tree1 and $C2_j$ in Tree2)
  - $C1_i = C2_j$ is translated into $C1_i \leftrightarrow C2_j$
  - $C1_i \sqsubseteq C2_j$ is translated into $C1_i \rightarrow C2_j$ (analogously for $\sqsupseteq$)
  - $C1_i \perp C2_j$ is translated into $\neg (C1_i \wedge C2_j)$

$$\text{Context} \rightarrow rel\ (C1_i, C2_j)$$

- A propositional formula is valid iff its negation is unsatisfiable
- SAT deciders are sound and complete…

# Step 4: cont'd (1)

```
1.  i, j, N1, N2: int;
2.  context, goal: wff;
3.  n1, n2: node;
4.  T1, T2: tree of (node);
5.  relation = {=,⊑ ,⊒ , ⊥ };
6.  ClabMatrix(N1, N2), CnodMatrix(N1, N2), relation: relation
7.      function mkCnodMatrix(T1, T2, ClabMatrix) {
8.          for (i = 0; i < N1; i++) do
9.              for (j = 0; j < N2; j++) do
10.                 CnodMatrix(i, j):=NodeMatch(T1(i),T2(j), ClabMatrix)}
11.     function NodeMatch(n1, n2, ClabMatrix) {
12.         context:=mkcontext(n1, n2, ClabMatrix, context);
13.         foreach (relation in  < =, ⊑, ⊒, ⊥ >) do {
14.             goal:= w2r(mkwff(relation, GetCnod(n1), GetCnod(n2));
15.             if VALID(mkwff(→, context, goal))
16.                 return relation;}
17.     return ⊓ ;}
```

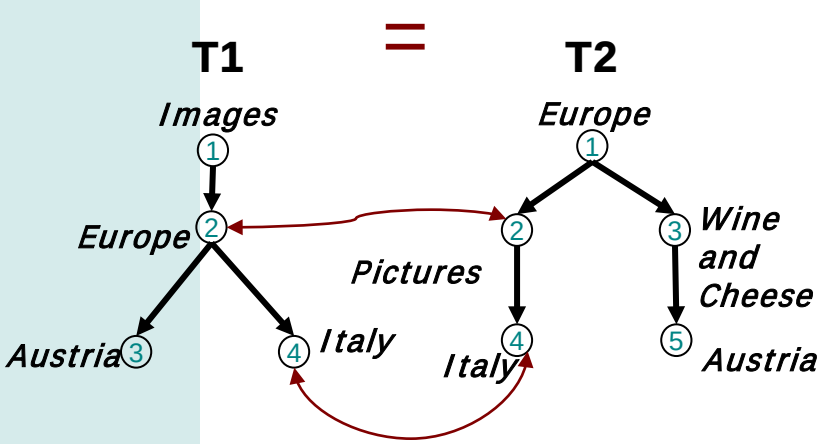
# Step 4: cont'd (2)

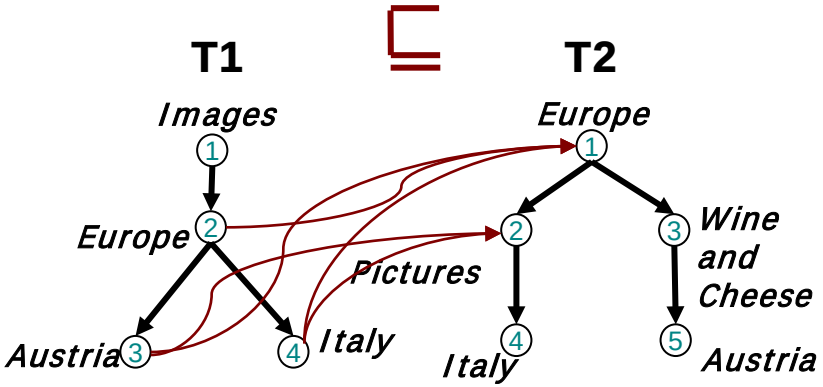
- Example. Suppose we want to check if $C1_2 = C2_2$

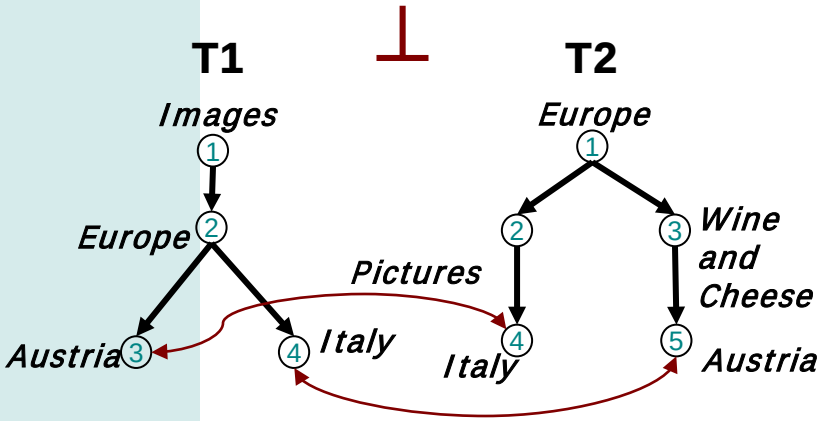
$$\underbrace{(C1_{Images} \leftrightarrow C2_{Pictures}) \wedge (C1_{Europe} \leftrightarrow C2_{Europe})}_{\text{Context}} \rightarrow \underbrace{(C1_2 \leftrightarrow C2_2)}_{\text{Goal}}$$

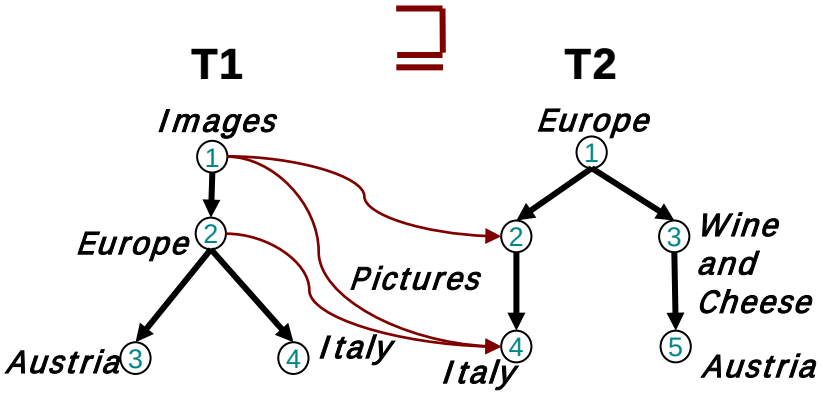
| T1 \ T2 | $C2_1$ | $C2_2$ | $C2_3$ | $C2_4$ | $C2_5$ |
|---|---|---|---|---|---|
| $C1_1$ | $\sqcap$ | $\sqsupseteq$ | $\sqcap$ | $\sqsupseteq$ | $\sqcap$ |
| $C1_2$ | $\sqsubseteq$ | $=$ | $\sqcap$ | $\sqsupseteq$ | $\sqcap$ |
| $C1_3$ | $\sqsubseteq$ | $\sqsubseteq$ | $\sqcap$ | $\bot$ | $\sqcap$ |
| $C1_4$ | $\sqsubseteq$ | $\sqsubseteq$ | $\sqcap$ | $=$ | $\bot$ |

# Step 4: cont'd (2)





# Element Level Semantic Matching

# Element level Semantic Matchers

- ***String based* matchers** have two labels as input and produce semantic relations exploiting string comparison techniques.
- ***Sense based* matchers** have two WordNet senses in input and produce semantic relations exploiting structural properties of WordNet hierarchies.
- ***Gloss based* matchers** have two WordNet senses as input and produce relations exploiting gloss comparison techniques.

# Element level Semantic Matchers: Overview

| Matcher name | Execution Order | Approximation level | Matcher type | Schema info |
|---|---|---|---|---|
| Prefix | 2 | 2 | String based | Labels |
| Suffix | 3 | 2 | | |
| Edit distance | 4 | 2 | | |
| Ngram | 5 | 2 | | |
| Text Corpus | 13 | 3 | | Labels + Corpus |
| WordNet | 1 | 1 | Sense based | WordNet senses |
| Hierarchy distance | 6 | 3 | | |
| WordNet Gloss | 7 | 3 | Gloss based | WordNet senses |
| Extended WordNet Gloss | 8 | 3 | | |
| Gloss Comparison | 9 | 3 | | |
| Extended Gloss Comparison | 10 | 3 | | |
| Semantic Gloss Comparison | 11 | 3 | | |
| Extended semantic gloss comparison | 12 | 3 | | |

# String based matchers: Prefix

**Prefix:** checks whether one input label starts with the other. It returns an equivalence relation in this case, and *Idk* otherwise.

| Source label | Target label | Relation | Time, ms |
|---|---|---|---|
| *net* | *network* | = | 0.00006 |
| *hot* | *hotel* | = | 0.00006 |
| *cat* | *core* | *Idk* | 0.00005 |

***Prefix*** is efficient in matching cognate words and similar acronyms (e.g., ***RDF*** and ***RDFS***) but often syntactic similarity does not imply semantic relatedness.

The matcher returns equality for ***hot*** and ***hotel*** which is wrong but it recognizes the right relations in the case of the pairs ***net***, ***network*** and ***cat***, ***core***.

# String based matchers: Edit distance

**Edit Distance:** calculates the edit distance measure between two labels. The calculation includes counting the number of the simple editing operations (delete, insert and replace) needed to convert one label into another. If the value exceeds a given threshold the equivalence relation is returned, otherwise, *Idk* is produced.

| Source label | Target label | Relation | Time, ms |
|---|---|---|---|
| *street* | *street1* | = | 0.019 |
| *proper* | *propel* | = | 0.016 |
| *owe* | *woe* | *Idk* | 0.007 |

# Sense based matchers: WordNet

**WordNet : return a relation which holds in WordNet between two input labels.**

**The relations provided by WordNet are converted to semantic relations according to the following rules:**

- **A ⊆ B if A is a hyponym, meronym or troponym of B;**
- **A ⊇ B if A is a hypernym or holonym of B;**
- **A = B if they are connected by synonymy relation or they belong to one synset;**
- **A ⊥ B if they are connected by antonymy relation or they are the siblings in the *part of* hierarchy**

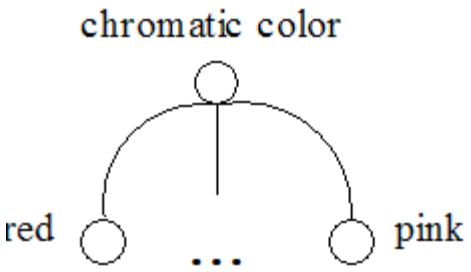
| Source label | Target label | Relation | Time, ms |
|---|---|---|---|
| *car* | *minivan* | ⊇ | 2,3 |
| *car* | *auto* | = | 0.6 |
| *tail* | *dog* | ⊆ | 0.2 |
| *red* | *pink* | *Idk* | 0.4 |

# Sense based matchers: Hierarchy distance

**Hierarchy distance:** returns the equivalence relation if the distance between two input senses in a WordNet hierarchy is less than a given threshold value and *Idk* otherwise.

| Source label | Target label | Relation | Time, ms |
|---|---|---|---|
| *red* | *pink* | = | 0.159 |
| *catalog* | *classification* | *Idk* | 0.203 |

There is no direct relation between ***red*** and ***pink*** *in WordNet*. However, the distance between these concepts is 2 (1 more general link and 1 less general). Thus, we can infer that ***red*** and ***pink*** are close in their meaning and return the equivalence relation.



# Gloss based matchers: WordNet gloss

**WordNet gloss:** compares the labels of the first input sense with the WordNet gloss of the second.

| Source label | Target label | Relation | Time, ms |
|---|---|---|---|
| *hound* | *dog* | $\subseteq$ | 0.031 |
| *hound* | *ear* | $\subseteq$ | 0.014 |
| *dog* | *cat* | *Idk* | 0.033 |

*Hound* is *any of several breeds of* ***dog*** *used for hunting typically having large drooping* ***ears.*** *Hound* is described through the specification of the more general concept ***dog.***

# Gloss based matchers: Gloss comparison

**Gloss comparison:** The number of the same words occurring in the two input glosses increases the similarity value. The equivalence relation is returned if the resulting similarity value exceeds a given threshold.

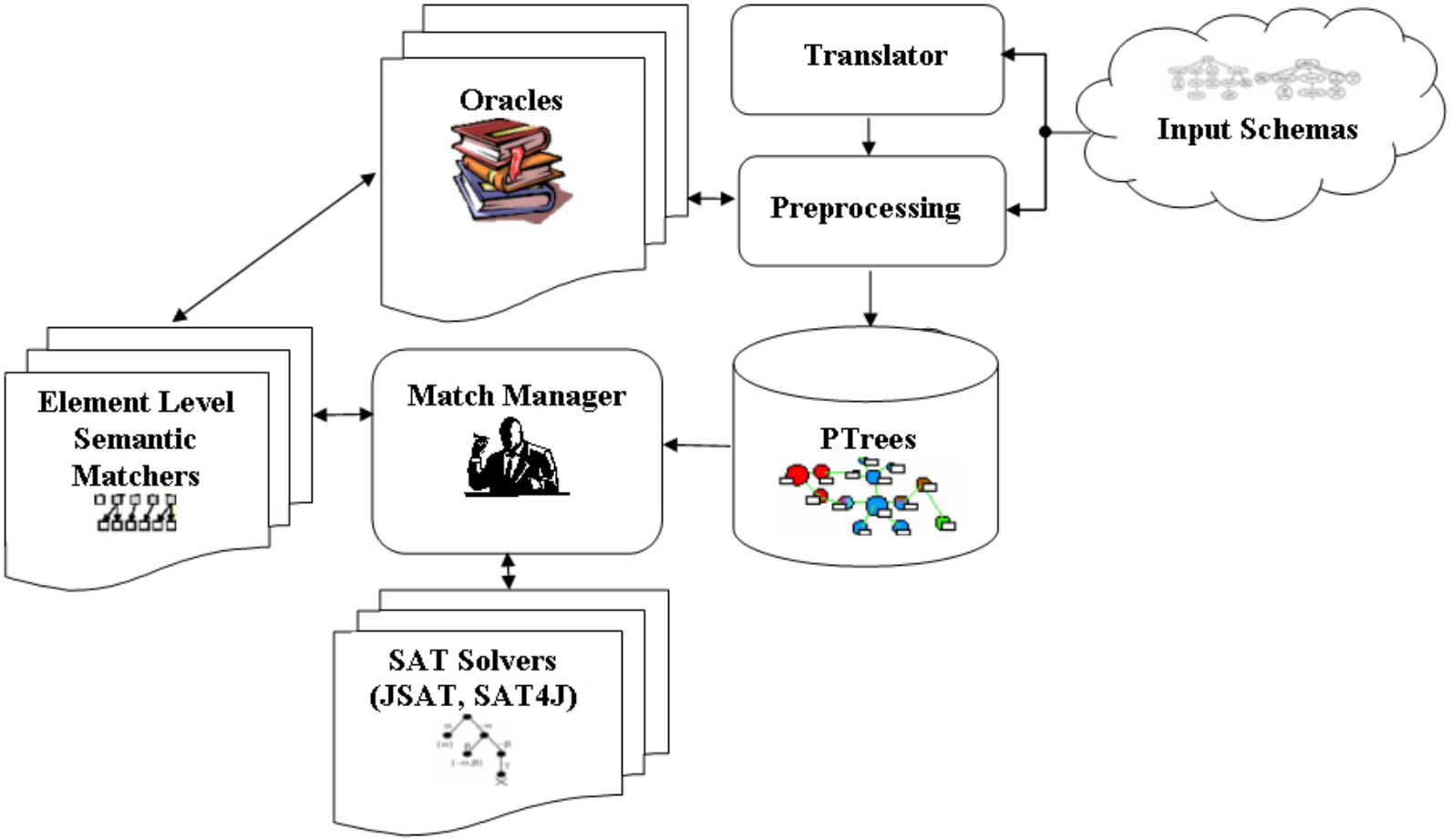
| Source label | Target label | Relation | Time, ms |
|---|---|---|---|
| *Afghan hound* | *Maltese dog* | = | 0,074 |
| *dog* | *cat* | *Idk* | 0,019 |

- *Maltese dog is a **breed** of toy dogs having a **long** straight **silky** white **coat***
- *Afghan hound is a tall graceful **breed** of hound with a **long silky coat**; native to the Near East*



# The S-Match System

# S-Match: Logical Level




**NOTE:** Current version of **S-Match** is a rationalized re-implementation of the **CTXmatch** system with a few added functionalities



# A Comparative Evaluation

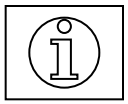
# Testing Methodology

## Matching systems

- S-Match vs. Cupid, COMA and SF as implemented in Rondo

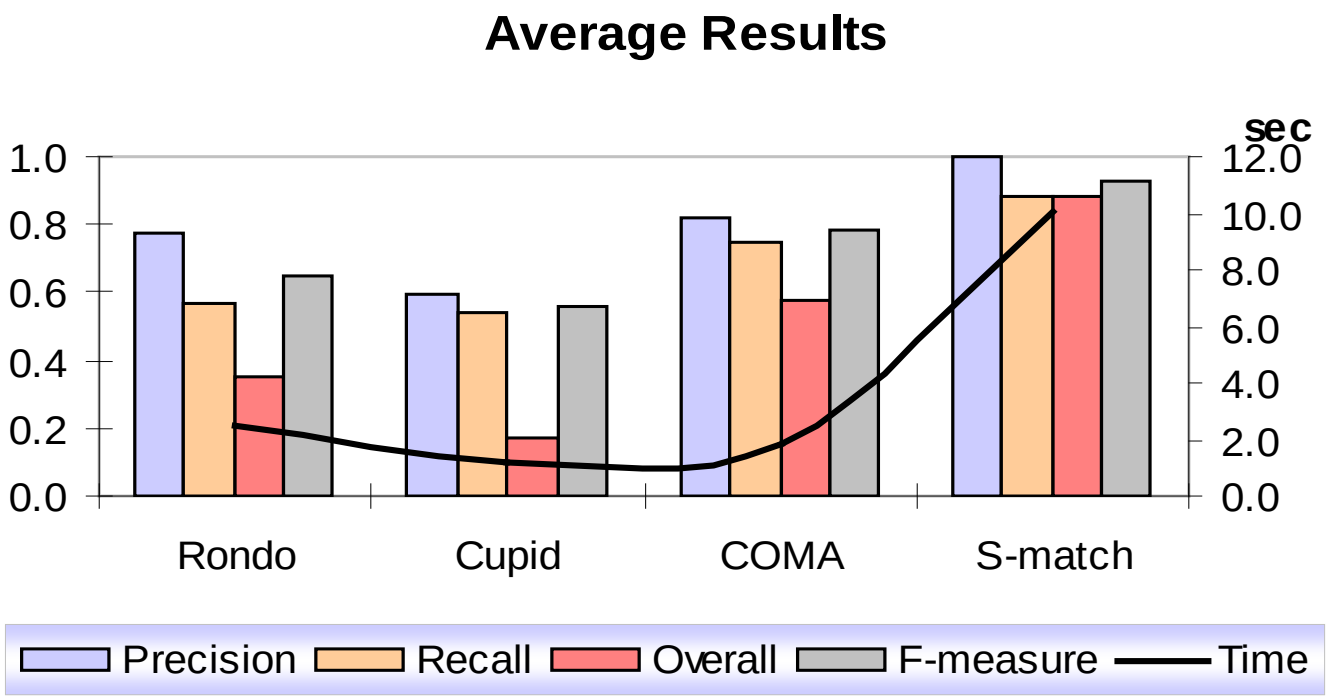
## Measuring match quality

- Expert mappings are inherently subjective
- Two degrees of freedom
  - Directionality
  - Use of Oracles
- Indicators
  - Precision, [0,1]
  - Recall, [0,1]
  - Overall, [-1,1]
  - F-measure, [0,1]
  - Time, sec.



# Preliminary Experimental Results

- Three experiments, test cases from different domains
- Some characteristics of test cases: #nodes 4-39, depth 2-3
- PC: PIV 1,7Ghz; 256Mb. RAM; Win XP





# Future Work

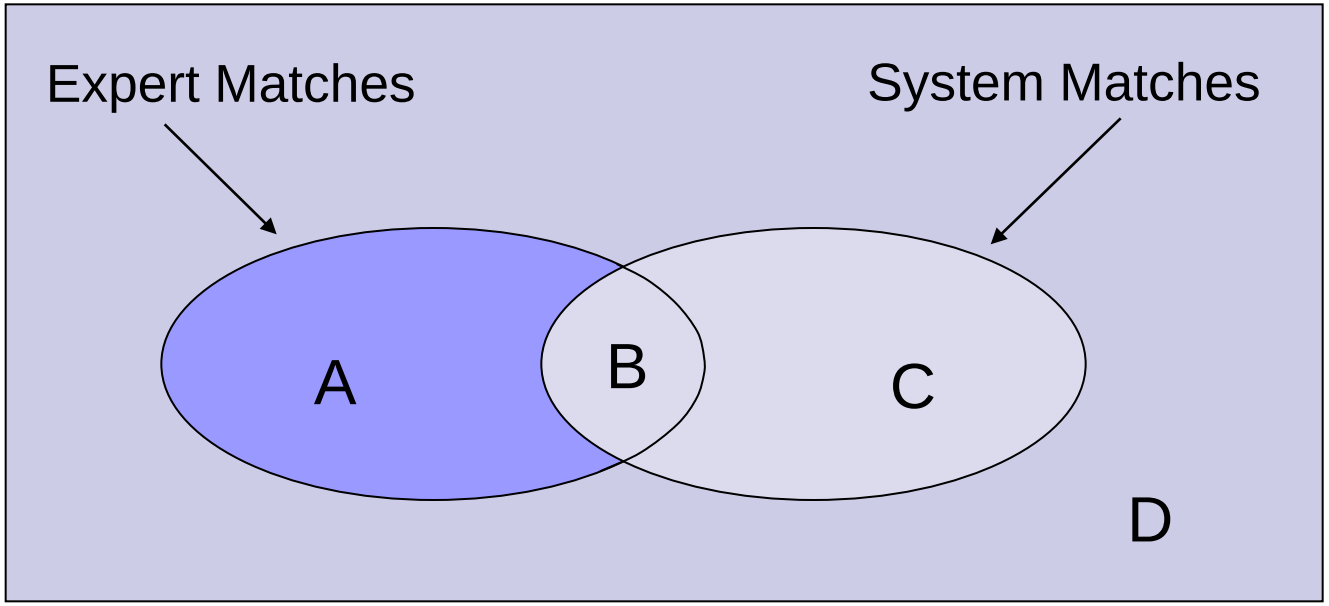
- Extend the semantic matching algorithm for computing mappings between graphs
- Develop iterative semantic matching
- Elaborate results filtering strategies according to the binding strength of the resulting mappings
- Efficient semantic matching
- Robust semantic matching
- Do throught testing of the system (small or big ontologies)

# References


- Project website - ACCORD: http://www.dit.unitn.it/~accord/
- F. Giunchiglia, P.Shvaiko, M. Yatskevich: S-Match: an algorithm and an implementation of semantic matching. In Proceedings of *ESWS'04*.
- F. Giunchiglia, P.Shvaiko: Semantic matching. In The Knowledge Engineering Review journal, 18(3):265-280, 2004. Short versions in Proceedings of *SI* workshop at *ISWC'03* and *ODS* workshop at *IJCAI'03*.
- P. Bouquet, L. Serafini, S. Zanobini: Semantic coordination: a new approach and an application. In Proceedings of *ISWC'03*.
- F. Giunchiglia, I. Zaihrayeu: Making peer databases interact – a vision for an architecture supporting data coordination. In Proceedings of *CIA'02.*
- C. Ghidini, F. Giunchiglia: Local models semantics, or contextual reasoning = locality + compatibility. Artificial Intelligence journal, 127(3):221-259, 2001.




# Thank you!



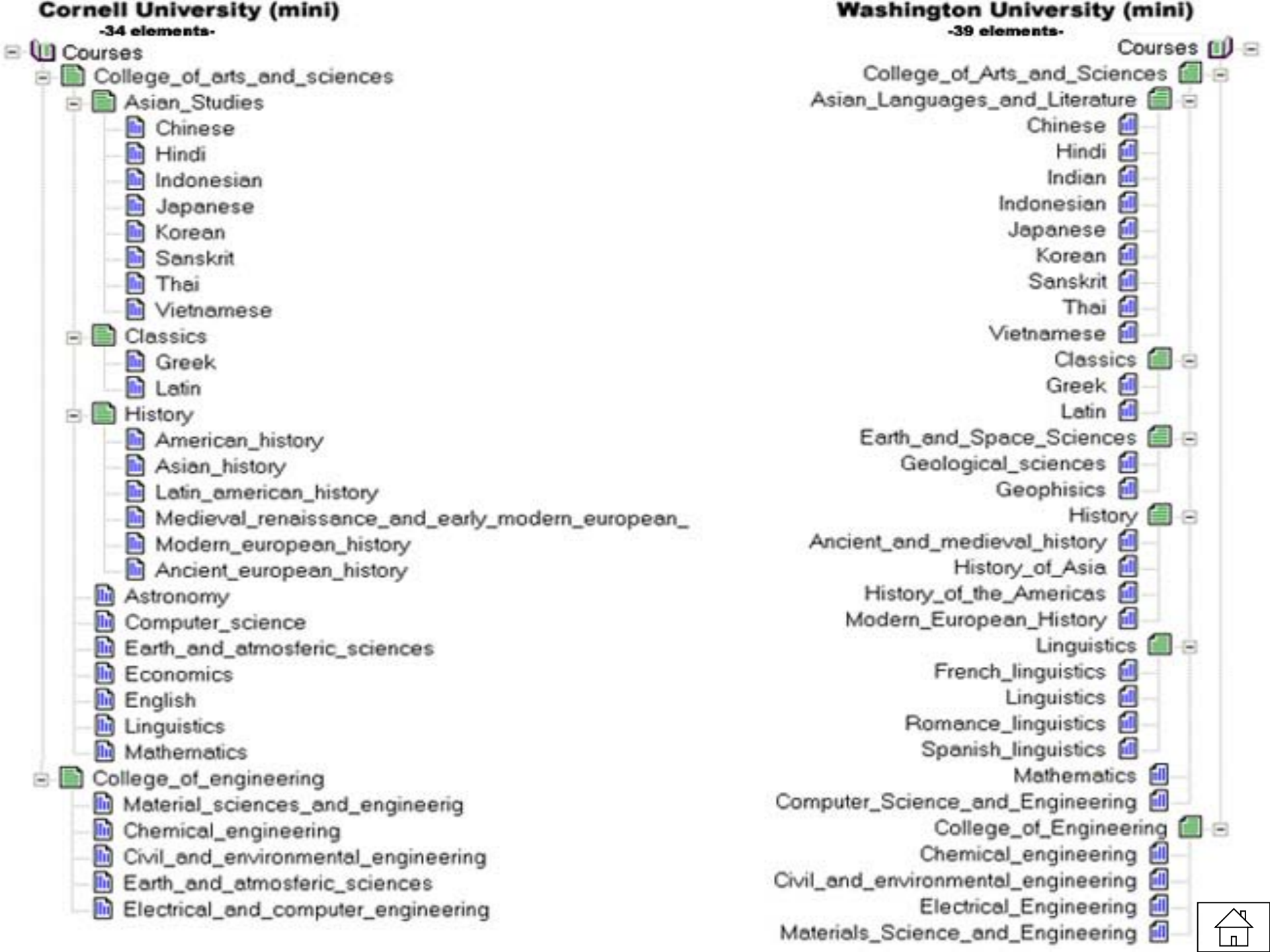

- A – False negatives
- B – True positives
- C – False positives
- D – True negatives

$$\text{Precision} = \frac{|B|}{|B| + |C|};$$

$$\text{Recall} = \frac{|B|}{|A| + |B|};$$

$$\text{Overall} = 1 - \frac{|A| + |C|}{|A| + |B|} = \frac{|B| - |C|}{|A| + |B|} = \text{Recall} \times \left(2 - \frac{1}{\text{Precision}}\right);$$

$$\text{F - Measure} = \frac{2 \times |B|}{(|A| + |B|) + (|B| + |C|)} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$


Cornell University (mini)
-34 elements-
Courses
College_of_arts_and_sciences
Asian_Studies
Chinese
Hindi
Indonesian
Japanese
Korean
Sanskrit
Thai
Vietnamese
Classics
Greek
Latin
History
American_history
Asian_history
Latin_american_history
Medieval_renaissance_and_early_modern_european_
Modern_european_history
Ancient_european_history
Astronomy
Computer_science
Earth_and_atmosferic_sciences
Economics
English
Linguistics
Mathematics
College_of_engineering
Material_sciences_and_engineerig
Chemical_engineering
Civil_and_environmental_engineering
Earth_and_atmosferic_sciences
Electrical_and_computer_engineering
Washington University (mini)
-39 elements-
Courses
College_of_Arts_and_Sciences
Asian_Languages_and_Literature
Chinese
Hindi
Indian
Indonesian
Japanese
Korean
Sanskrit
Thai
Vietnamese
Classics
Greek
Latin
Earth_and_Space_Sciences
Geological_sciences
Geophisics
History
Ancient_and_medieval_history
History_of_Asia
History_of_the_Americas
Modern_European_History
Linguistics
French_linguistics
Linguistics
Romance_linguistics
Spanish_linguistics
Mathematics
Computer_Science_and_Engineering
College_of_Engineering
Chemical_engineering
Civil_and_environmental_engineering
Electrical_Engineering
Materials_Science_and_Engineering