DEFENCE R&D DÉFENSE

Measures of Similarity for Command and Control Situation Analysis

Eric Dorion
Alexandre Bergeron Guyard
DRDC Valcartier

June 2011
16th ICCRTS, Québec, Canada

Defence Research and Development Canada
Recherche et développement pour la défense Canada
Canada

Presentation Layout

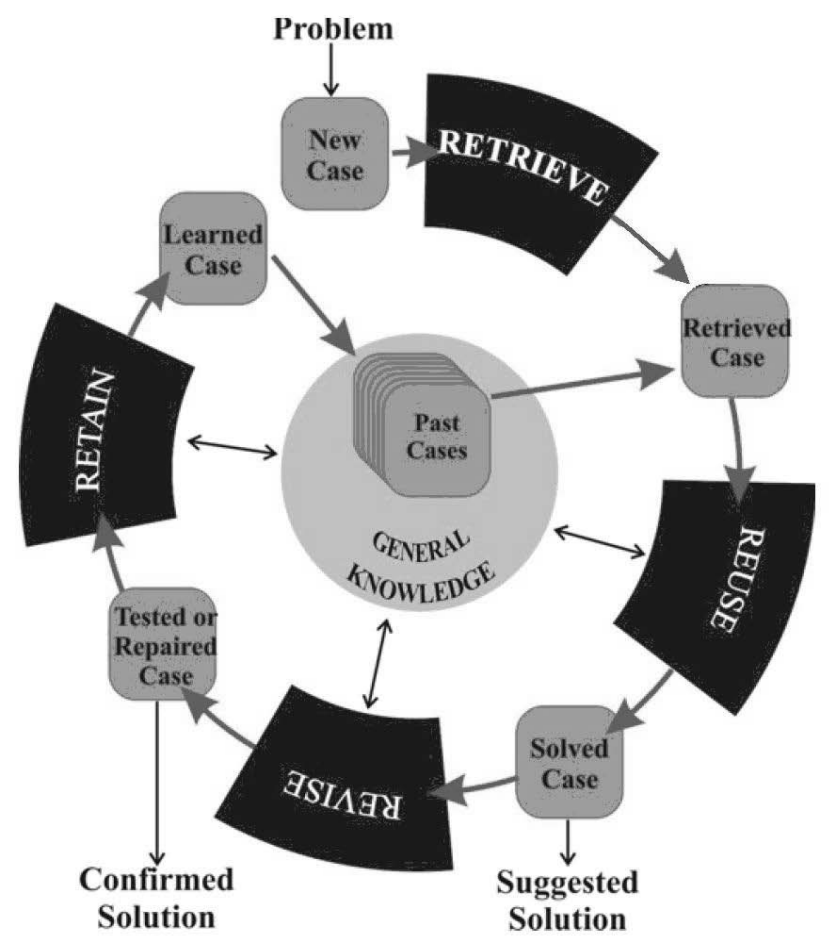
- Introduction
- Case-Based Reasoning for C2 Information Systems
  - CBR Challenge : Measuring Similarity
- Measures of Similarity
  - Geometry-based measures
  - Feature-based measures
  - Structure-based measures
  - Transformation-based measures
  - Information Content-based measures
- Choosing a Measure of Similarity for CBR
- Conclusion

Introduction

- Command and Control (C2) in Multinational Civil-Military Operations is a demanding environment
- Information Systems supporting this environment need adaptable automated reasoning capabilities
- Case-Based Reasoning offers a flexible approach to automated reasoning compared to formal logic approaches.
- CBR relies on the ability to establish *similarity* between unfolding situations (current cases) and known cases (from a case base)
- Measuring similarity is CBR's Achile's heel.
- Understanding measures of similarity is of prime importance to successfully apply CBR.

Case-Based Reasoning Basic Concepts

- The Case-Based Reasoning Cycle
  - Retrieve *similar* cases to the problem description
  - Reuse a solution suggested by a *similar* case
  - Revise or adapt that solution to better fit the new problem
  - Retain the new solution once it has been confirmed or validated



Case-Based Reasoning Basic Concepts

- The Case-Based Reasoning Challenges
  - A standard problem template must be produced in order to describe and organize problems in a way that will allow comparison
  - In order to retrieve a similar problem from the case base's problem space, there must be a way to *measure similarity* between problems

Measures of Similarity

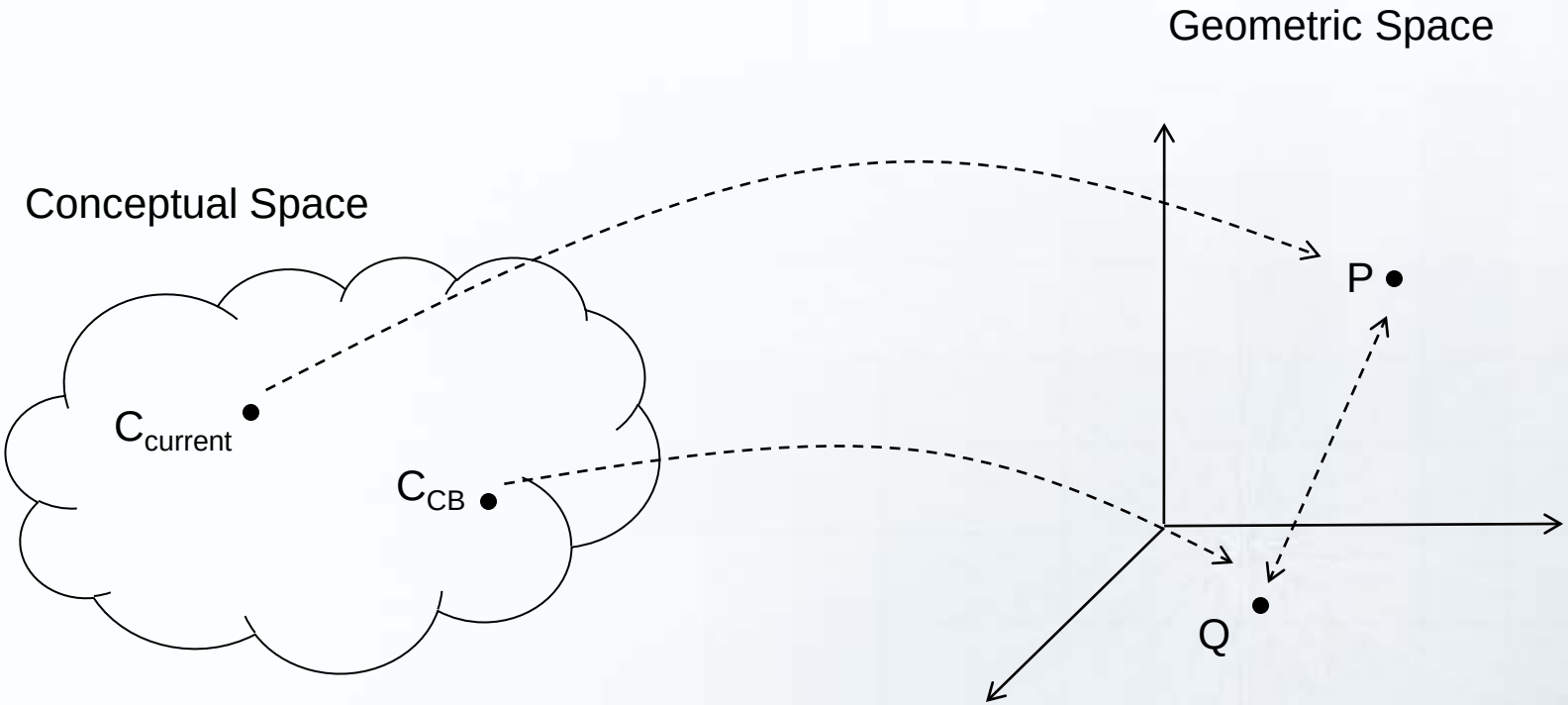
- Geometry-based Measures
- Feature-based Measures
- Structure-based Measures
- Transformation-based Measures
- Information Content-Based Measures

Measures of Similarity

- **Geometry-based Measures**
- Feature-based Measures
- Structure-based Measures
- Transformation-based Measures
- Information Content-Based Measures

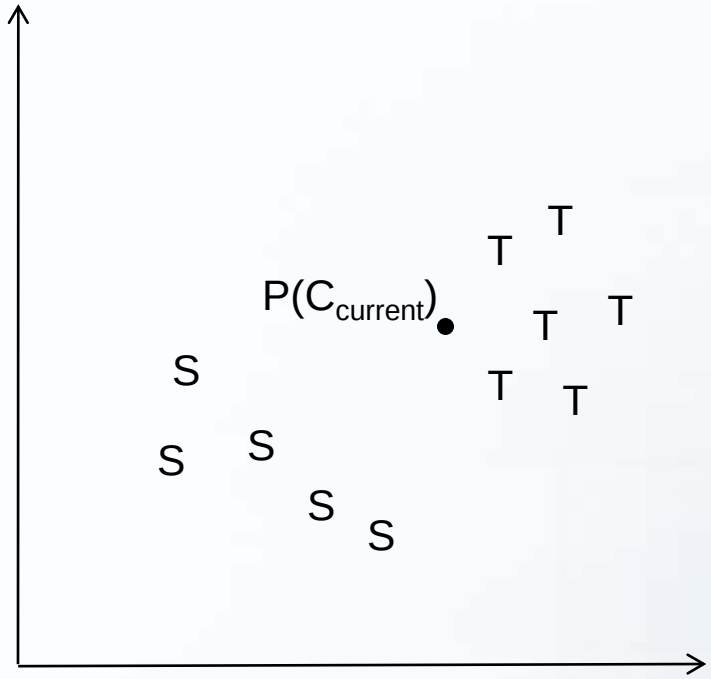
Measures of Similarity

- Geometry-based Measures



Measures of Similarity

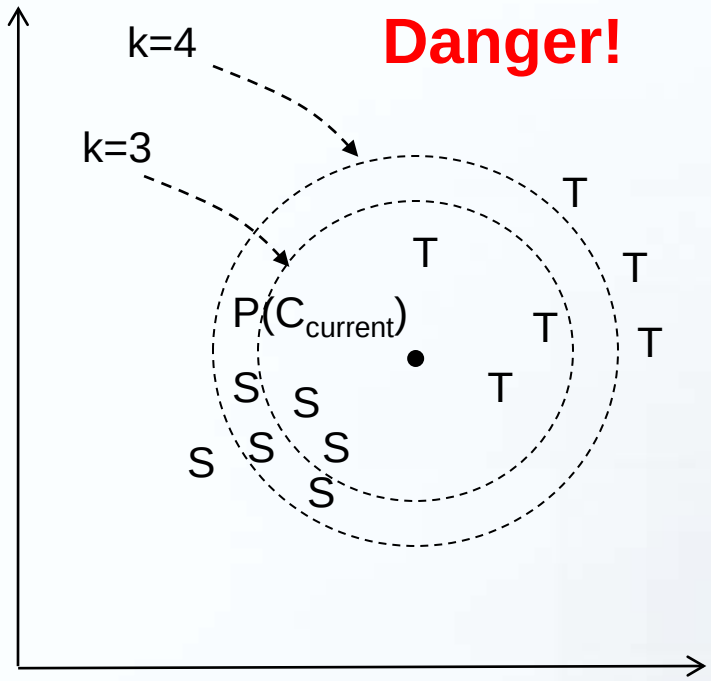
- Geometry-based Measure: K Nearest Neighbor Rule



$P(C_{current})$ is associated with Case T

Measures of Similarity

- Geometry-based Measure: K Nearest Neighbor Rule

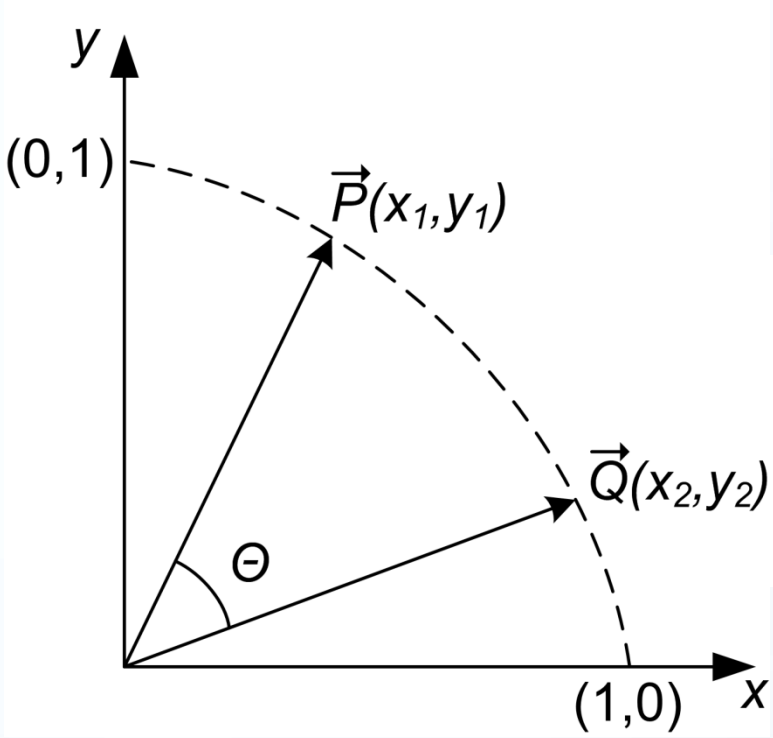


If k=3, then $P(C_{current})$ is associated with Case T.

If k=4, then $P(C_{current})$ is associated with Case S.

Measures of Similarity

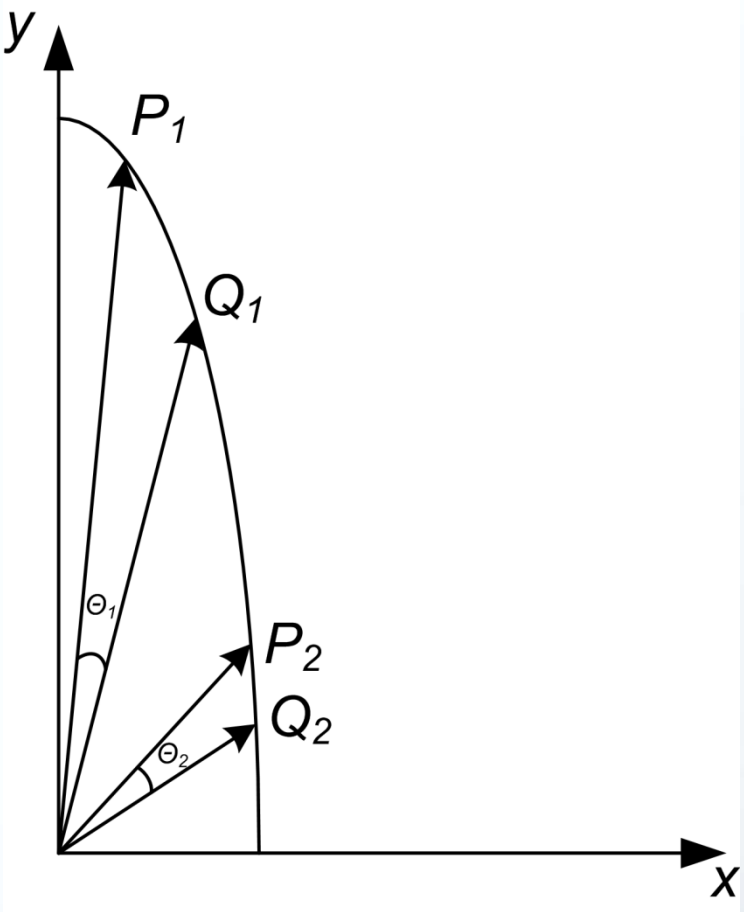
- Geometry-based Measure: Cosine Similarity



$$sim(C_{cur.}, C_{CB}) = \begin{cases} Identical, & \text{if } \cos(\vec{P}, \vec{Q}) = 1; \\ Independent, & \text{if } \cos(\vec{P}, \vec{Q}) = 0; \\ Opposite, & \text{if } \cos(\vec{P}, \vec{Q}) = -1; \\ (dis)similar, & \text{otherwise.} \end{cases}$$

Measures of Similarity

- Geometry-based Measure: Cosine Similarity (Limit?)



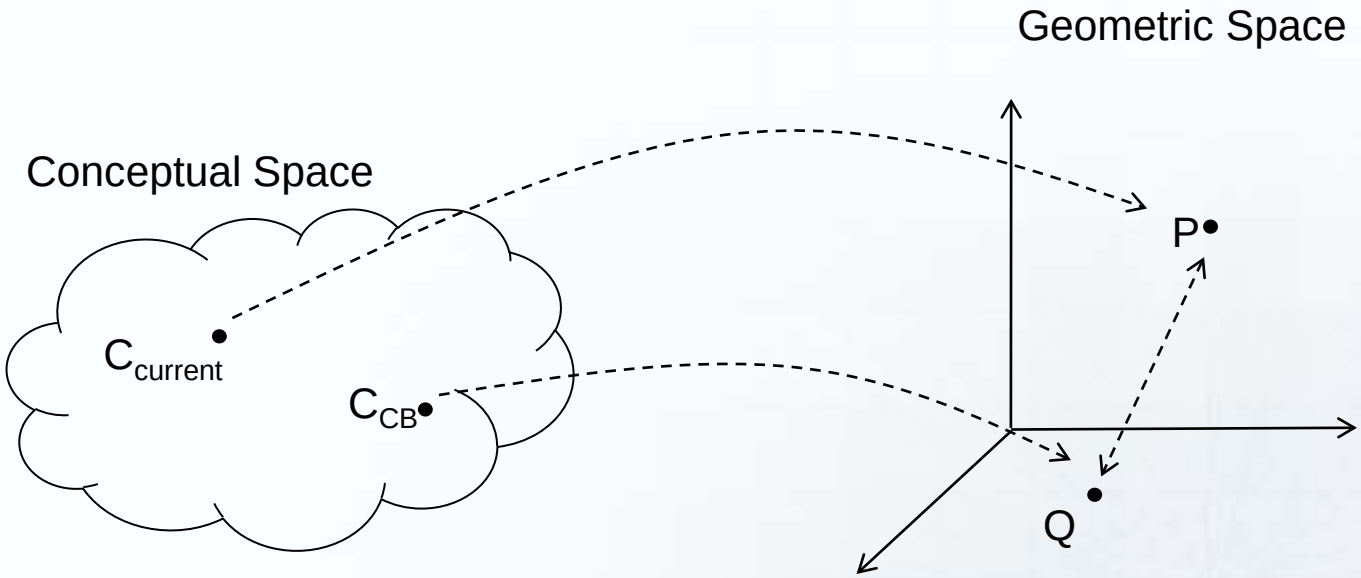
Feature *y* is more important then *x.*

$\theta_1 = \theta_2$ implies that $\cos(\theta_1) = \cos(\theta_2)$, but $dist(P_1,Q_1) > dist(P_2,Q_2)$.

What does that mean in the conceptual space?

Measures of Similarity

- Geometry-based Measure: Limits



Geometry Axioms

Amos Tversky 1977 showed examples where these axioms were violated.

$$d(x,y) \geq 0; \text{ non-negativity},$$

$$d(x,y) = 0 \text{ iff } x = y; \text{ identity},$$

$$d(x,y) = d(y,x); \text{ symmetry},$$

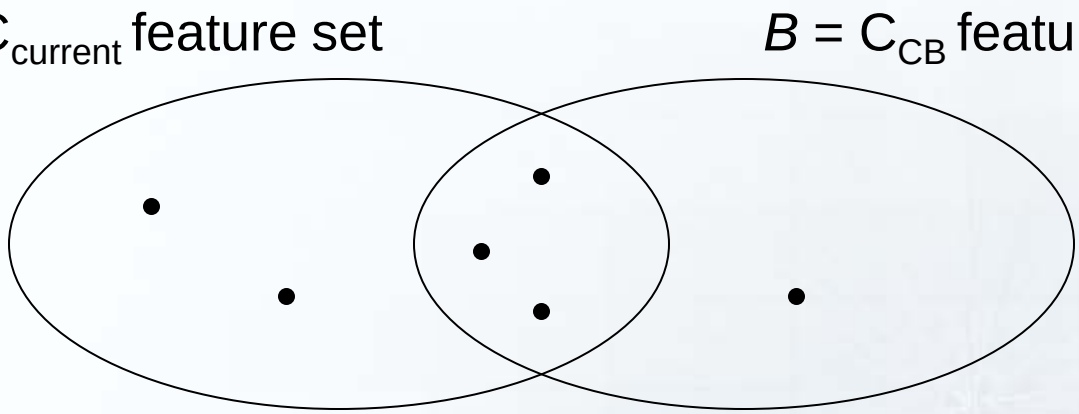
$$d(x,z) \leq d(x,y) + d(y,z); \text{ triangle inequality}$$

Measures of Similarity

- Geometry-based Measures
- **Feature-based Measures**
- Structure-based Measures
- Transformation-based Measures
- Information Content-Based Measures

Measures of Similarity

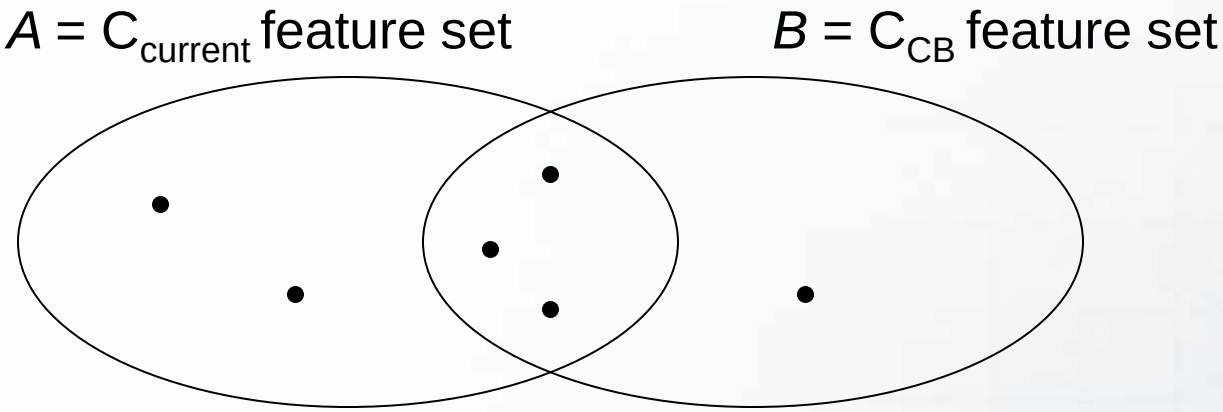
- Feature-based Measures



Tversky's index = $$S(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha \cdot |A - B| + \beta \cdot |B - A|}$$

Measures of Similarity

- Feature-based Measures: Limits



This similarity measure only depends on the feature *count*.

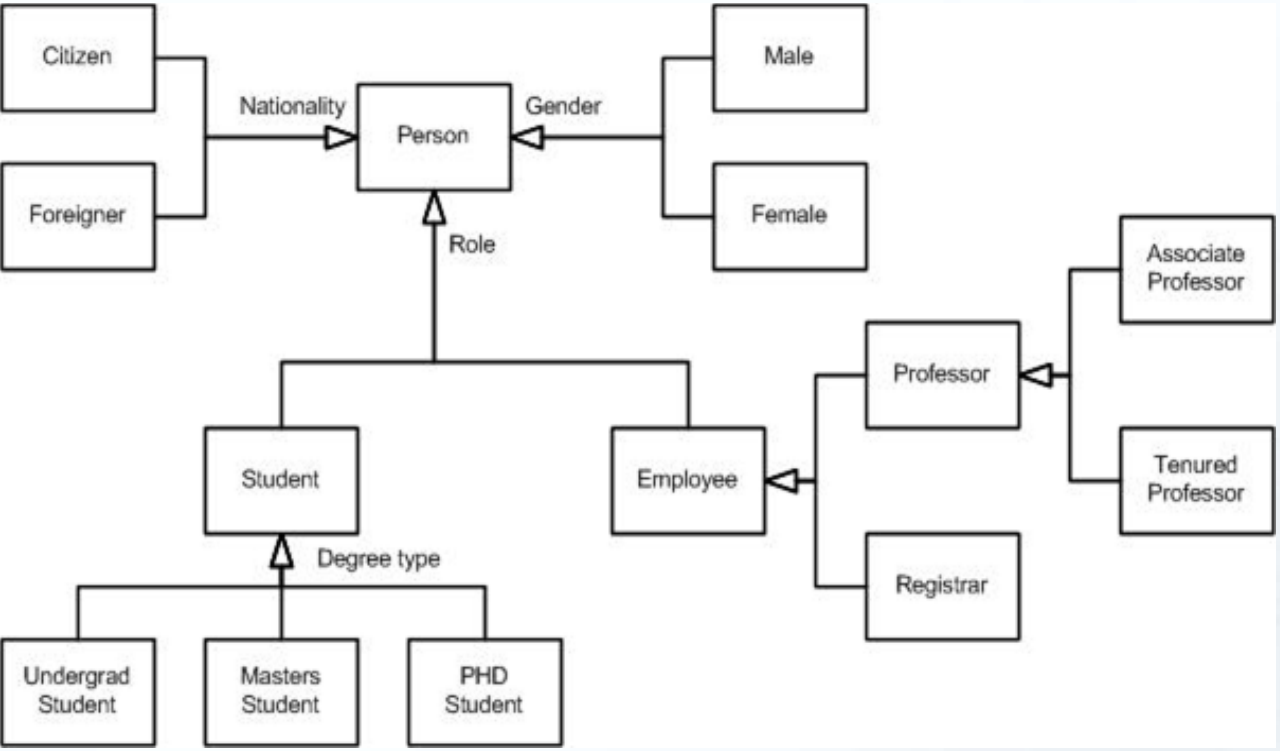
Does not take into account feature weighting.

Measures of Similarity

- Geometry-based Measures
- Feature-based Measures
- **Structure-based Measures**
- Transformation-based Measures
- Information Content-Based Measures

Measures of Similarity

- Structure-based Measures



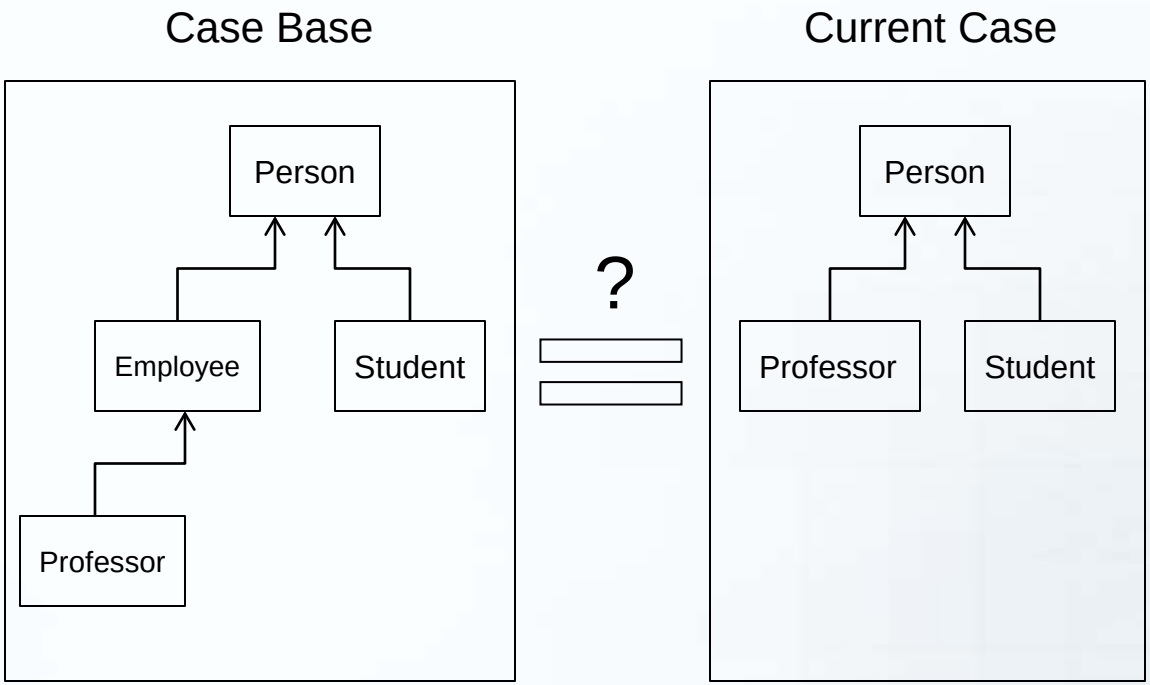
Sim(Male,Female)=1

Sim(PhD Student,Male)=2

Sim(Tenured Prof,Registrar)=2

Measures of Similarity

- Structure-based Measures: Limits



- Similarity value highly dependent on structure and semantic content
- Difficult to compare concepts of distinct structures

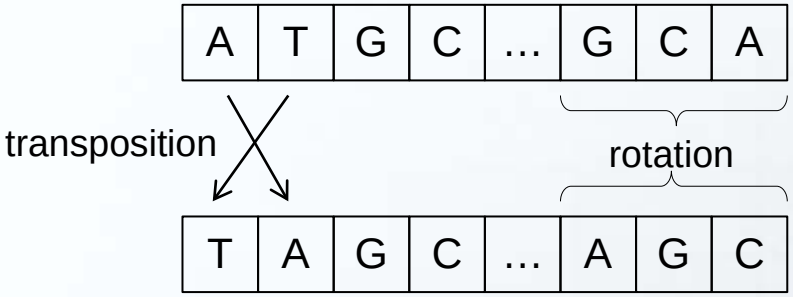
Measures of Similarity

- Geometry-based Measures
- Feature-based Measures
- Structure-based Measures
- **Transformation-based Measures**
- Information Content-Based Measures

Measures of Similarity

- Transformation-based Measures

  TBMs count how many allowable transforms are necessary to shape one pattern into the other.



  - Hamming: How many differences?
  - Levenshtein: How many *inserts, deletes* and *substitutes?*
  - Damerau–Levenshtein: Levenshtein + *transposition*
  - etc.

Measures of Similarity

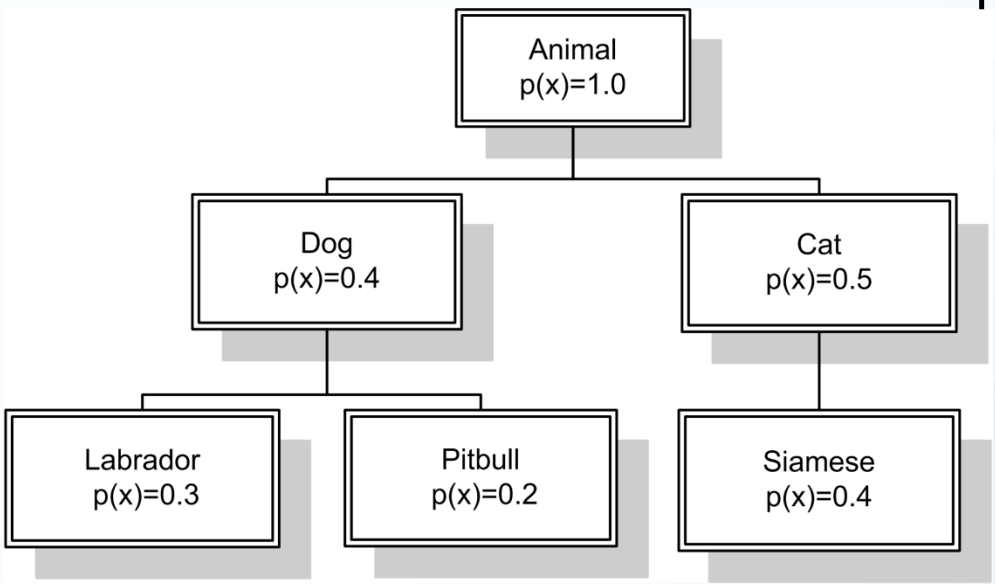
- Transformation-based Measures: Limits
  - Computes differences between *strings* (DNA, bitstreams, speech flow, etc.)
  - Cases representation complexity
  - Cases (between the Case base and the current case) may present too many differences (topological, semantics, context) for the measure to make sense
  - Difficult to establish what are the allowable and relevant transformations for between cases.

Measures of Similarity

- Geometry-based Measures
- Feature-based Measures
- Structure-based measures
- Transformation-based Measures
- **Information Content-Based Measures**

Measures of Similarity

- Information Content-Based Measures



Resnik 1999

- Based on Information Theory
- The similarity between 2 concepts within a taxonomy corresponds to the *information content* of the closest parent concept.
- A probability of *occurrence* is associated to every concept in the hierarchy.

*Sim(*Pitbull, Labrador) = Information content of *Dog* = - log(0.4) = 0.92

Sim(Siamese, Dog) = Information content of *Animal* = - log(1.0) = 0

Choosing a Similarity Measure for CBR

- Case representation and feature selection is a hard problem because the application domain is complex (C2 in Multinational Civil-Military Operations)
- A mix of Geometry-based and Feature-based measures have been used in DRDC Valcartier.
  - Lesson learned: Must take into account the military operation. (Scalability?)
- Structure-based measures' reliability depend on case representation, which varies a lot (many different standards).
- Transformation-based measures suffer from being too "local" (string-based). Difficult to see how it could apply to complex cases.
- Information content-based measures are under development. Resnik's approach too restrictive (what about Entropy?)

Conclusion

- The problem of characterizing Command and Control (C2) in Multinational Civil-Military Operations: What are the most relevant features?
- C2ISs using Case-Based Reasoning seems to be promising, but is the solution scalable and shareable to the Multinational Civil-Military Operational environment?
- CBR requires the ability to establish *similarity*, which is a cognitive process. Are we there yet? More research? Most certainly!


DEFENCE
R&D
DÉFENSE