16th ICCRTS

“Collective C2 in Multinational Civil-Military Operations”

Paper Title:
Measures of Similarity for Command and Control Situation Analysis

Topic:
Topic 4: Information and Knowledge Exploitation

Authors:
Eric Dorion
Alexandre Bergeron Guyard

Defense Research and Development Canada – Valcartier
2459, Pie XI North Quebec, Quebec, Canada, G3J 1X5
{Eric.Dorion, Alexandre.Bergeron-Guyard}@drdc-rddc.gc.ca

Measures of Similarity for Command and Control Situation Analysis

Eric Dorion
Alexandre Bergeron Guyard
Defense Research and Development Canada – Valcartier
2459, Pie XI North, Quebec, Quebec, Canada, G3J 1X5
{Eric.Dorion, Alexandre.Bergeron-Guyard}@drdc-rddc.gc.ca

Abstract

Collective Command and Control (C2) in Multinational Civil-Military Operations pose stringent requirements on C2 Information Systems (C2ISs) interoperability as well as the higher level automated reasoning processes. One such process, *Case-Based Reasoning* (CBR), depends on proper case formal representations and on the ability to assess similarity between an unfolding case (e.g. C2 situation) and known cases *a priori* in the case base. This paper explores the main classes of similarity metrics and reflects on the most desirable features of an ideal similarity measure in support of C2 situation analysis.

1 Introduction

The need to support *Collective Command and Control (C2) in Multinational Civil-Military Operations* pose stringent requirements on C2 Information Systems (C2ISs) interoperability as well as the higher level automated reasoning processes. While the dynamics of military coalitions are known to be complex in nature[1], some issues with supporting automated reasoning in C2ISs need to be considered in order to achieve effective support to C2. One such method of formal reasoning in computer science, Case-Based Reasoning (CBR), has drawn attention over the years as a promising tool to support the sense-making process[2]. Applied research projects in DRDC Valcartier are currently using CBR techniques to support C2 situation analysis and maritime anomaly detection[3]. However, the potential impact of some sensitive aspects of CBR on C2 support has yet to be taken into account in these initiatives. Notably, CBR is highly dependant on cases formal representation (knowledge representation) and the establishment of similarity between the problem under study and the problems base[4]. This impacts the case retrieval process.

While many research initiatives in Valcartier addressed knowledge representation from a broad perspective [5, 6, 3], or more particularly to CBR [7] none have tackled, as a research focus, what would be the most sensible way to establish similarity for C2 CBR. This paper aims at bringing into focus how CBR similarity is established,

exposing the many different strategies that currently exists, highlighting their strengths and weaknesses and finally discussing the characteristics of a similarity metric suitable to C2 situation analysis.

The paper is structured as follows: Section 2 explains CBR basic concepts, section 3 describes some of the most prominent measures of similarity, section 4 brings back into consideration how CBR paired with an appropriate measure of similarity can support C2. and section 5 brings concluding remarks.

2 Case-based Reasoning Basic Concepts

A case-based reasoner solves current problems by using or adapting prior solutions to old problems[8]. The general idea is to emulate the human reasoning process that relies on past experiences to solve new problems, reusing past solutions. The classical example considers how doctors establish diagnoses and treatments based on the successful matching of apparent symptoms to known diseases, thus relying on past experience of thousands of *similar* cases[9]. It is hypothesized in this approach that new cases (sets of symptoms) will bear sufficient similarity with known cases (diseases with typical symptoms) to allow an appropriate matching. Case-Based Reasoning (CBR) systems thus require a storage component, the *case base*, and some means to match unfolding cases to known cases in the case base. Typically, a known case, described by some knowledge representation means is paired with a specific solution to address the problem at hand.

2.1 The Case-based Reasoning Cycle

Case-based Reasoning can also be explained by considering its CBR *cycle* (Figure 1), termed *the four Rs* in Aamodt and Plaza[2]:

1. Retrieve similar cases to the problem description,
2. Reuse a solution suggested by a similar case,
3. Revise or adapt that solution to better fit the new problem,
4. Retain the new solution once it has been confirmed or validated.

Any new problem is first compared to other cases present in the case base in order to retrieve one or more similar cases. The associated solutions are proposed for reuse to the new problem. With the revise process, the solution is tested for success and validated by an expert. At this point, the solution may have to be adapted in order to solve the problem more efficiently. Finally, in the retain process, the validated, adapted solution and the new problem are added to the case-base for future reuse.

2.2 CBR Challenges

Achieving CBR and applying the four Rs leads to a series of tasks and challenges, among which:

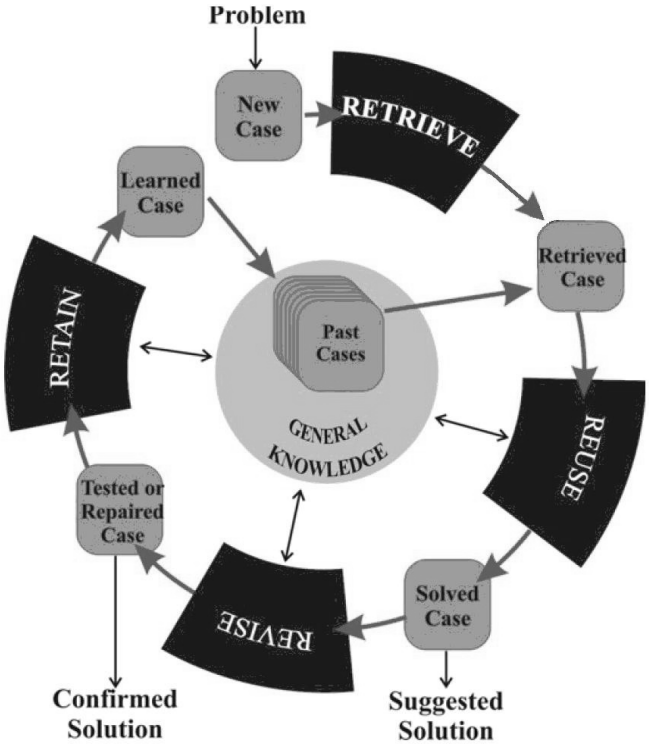

Figure 1: The CBR Reasoning Cycle

1. A standard problem template must be produced in order to describe and organize problems in a way that will allow comparison.

2. In order to retrieve a similar problem from the case base's problem space, there must be a way to measure similarity between problems.

The challenge in building an appropriate problem template lies in carefully selecting attributes that accurately describe a given situation. Selecting too few attributes might result in not being able to properly describe a number of emerging problems. A lack of relevant attributes may also lead to difficulties in building an efficient similarity measure. On the other hand, selecting too many attributes would make populating problem templates tedious. A large amount of attributes would also possibly lead to having many seldom used, less relevant attributes, which would add noise, and potential imprecision, to the similarity measure. In order to retrieve similar problems from the case base, there must be a way to measure the similarity between problems. There are many ways to measure similarity. Selecting a proper and efficient similarity measure is the central topic of this paper, and is discussed in greater details in the following sections.

3 Measures of Similarity

CBR relies on the establishment of similarity between a current case under assessment and a case base. CBR, in effect, emulates part of the human reasoning process where past experience is stored in memory for later retrieval. In cognitive psychology and later in computer science, many research efforts have been deployed to discover the most appropriate similarity measuring tool[1] that will emulate the human recollection of past experience (the case base).

This section presents these measuring tools grouped into five main categories: Geometry-based, feature-based, structure-based, transformation-based and Information content-based measures. There is no specific consensus on this classification (exemplified when comparing Cunningham [4] and Goldstone *et al* [10]) that would support the building of a similarity measure ontology, but it still helps in understanding the major trends and efforts since the 1970s.

3.1 Geometry-based Measures

Geometry-based measures assess similarity as a function of *distance* between objects in a n-dimensional euclidian space. That is, the smaller the distance between objects, the greater is the similarity between them. Applied to CBR, cases from the case base would be retrieved based on the *distance* with the new case (from the unfolding situation). Each of the n dimensions of the euclidian space correspond to specific features of importance to the CBR system. In this metric approach, the similarity between a current case $C_{current}$ and another one C_{CB} from the case base is first established by porting those into a metric space where they will be represented by points. Formally,

$$C_{current} \rightarrow P = (x_1, \ldots, x_n), C_{CB} \rightarrow Q = (y_1, \ldots, y_n) \tag{1}$$

where n is the number of relevant features of the CBR problem space (e.g. Situation analysis, threat evaluation, anomaly detection, etc.) Similarity is then modeled as an inverse function of *distance* between the geometric points counterparts:

$$sim(C_{current}, C_{CB}) \models f \propto 1/d : P \times Q \rightarrow \mathbb{R} \tag{2}$$

In geometry, distance d can be established in different fashions, notably with Minkowski's distance function family

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{3}$$

Here, p determines the specific distance function. For example, $p = 1$ yields the Manhattan distance (city-block distance), $p = 2$ yields the Euclidian distance and $p = \infty$ yields the Chebyshev distance.

In CBR a classical algorithm that makes use of geometric distances is the *K Nearest Neighbor Rule*.

[1] In this paper the term *metric* will be used for the measuring tools used in Geometry. From a pure ontological perspective, a metric implies Geometry and thus precludes those measuring tools that are not using that theory.

3.1.1 K Nearest Neighbor Rule

The K Nearest Neighbor Rule (k-NNR) is a method of classification often used in CBR. Its typical implementation is to assess the closeness of an unclassified object to k labeled (classified or training) objects in a n-dimensional metric space. A decision procedure assigns the unknown object to a specific class, based on the smallest average euclidian distance with the k neighbors. Like CBR, k-NNR is often viewed as a *lazy-learning* method since it defers its classification function until it is queried. It therefore has less stringent requirements on *a priori* training but necessitates more computation each time it is queried and thus requires more storage.

k-NNR's performance is known to be sensitive to feature weighting, i.e. the decision procedure for classification is blurred if a specific feature is much more conspicuous than the others (two order of magnitude in feature scaling can produce dramatic results). If normalization of the most salient features is an option to consider for better performance of the algorithm, then the limiting factor becomes how faithful the feature representation is. Indeed, if a specific feature in the conceptual (real-world) space is known to be more conspicuous, then its the process of normalization to ease the k-NNR process inevitably brings distortion. Also, as the number of features increases, k-NNR's performance tends to decrease as the distance between unclassified data and other classes also decreases, resulting in higher uncertainty in the classification[11].

3.1.2 Cosine Similarity

The cosine similarity metric makes use of the same isomorphism expressed by Equation 1. From the two vectors $\overrightarrow{P} = \langle (0,0), P \rangle$ and $\overrightarrow{Q} = \langle (0,0), Q \rangle$, the cosine of the angle they form is calculated. The resulting value is then reinterpreted in the conceptual space as a degree of similarity. Formally,

$$sim(C_{current}, C_{CB}) \models f \propto \cos(\overrightarrow{P}, \overrightarrow{Q}). \tag{4}$$

As the cosine value varies from -1 to 1, Three special meanings of similarity are often attributed to values -1, 0 and 1, such that:

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

3.1.3 Applicability of Geometry-based Measures to CBR

Geometry-based metrics constitute a strong body of measuring tools that is still actively being researched[12, 13]. Geometry provides in effect many efficient tools to assess the *distance* between any two constructs embedded in the metric space. The real challenge concerns the isomorphism that must be established between the CBR cases and their geometric counterparts. That is, how can we ensure that a *distance* effectively models the notion of similarity between cases as stated in Equation 2? Also, since the

geometric model strongly leans to the establishment of a number of orthogonal dimensions to establish the metric space, what can be said about those non-orthogonal case features that are still determinant in the decision-making process (the *Retrieve* phase of the CBR cycle)?

Perhaps one of the most serious blow against geometric models of similarity came from Amos Tversky in 1977[14]. By considering the axiomatic foundation of Geometry, namely,

1. $d(x,y) \geq 0$; non-negativity,
2. $d(x,y) = 0$ iff $x = y$; identity,
3. $d(x,y) = d(y,x)$; symmetry,
4. $d(x,z) \leq d(x,y) + d(y,z)$; triangle inequality,

Tversky showed how the establishment of similarity in psycho-cognitive experiments seemed not to follow geometry axioms, notably on identity, symmetry and triangle inequality. This in effect questioned the validity of modeling similarity with Geometry from its very axiomatic foundation. Tversky then proposed his own measure of similarity that spawned the *feature-based* family of similarity measures.

3.2 Feature-based Measures

To alleviate the apparent inadequacies of geometry-based metrics, Tversky proposed a metric that makes use of the number of similar and dissimilar features between objects. Tversky's argument is based on the assumption that, in comparing two concepts A and B for similarity, the more *features* they share, the more similar they are. He also reflected that the more features distinguish them, then the more dissimilar they are. His measuring tool

$$S(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha \cdot |A - B| + \beta \cdot |B - A|} \tag{5}$$

computes the similarity between A and B based on the number of features they share and do not share. The weighting factors α and β take into account possible asymmetric aspects of similarity between A and B. In considering, for example, the similarity between a *person* and his *portrait*, it is acceptable (cognitively) to say that this portrait *resembles* (is similar to) that person but not that the person resembles the portrait. Tversky's index can consider asymmetry where geometric models axiomatically cannot.

3.2.1 Jaccard Index and Dice's Coefficient

Setting $\alpha = \beta = 1$ in Equation 5 yields the Jaccard index,

$$S_J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

which has been used in Biology to establish resemblance between two community species. Dice's coefficient is linearly proportional to the Jaccard index, but gives twice more weight on the shared attributes.

$$S_D(A,B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2 \cdot S_J(A,B)}{1 + S_J(A,B)} \tag{7}$$

Tversky's index, Jaccard index and Dice's coefficient are essentially mathematically equivalent. They have been used for different purposes since their inception. They stress out however the fact that feature weighting is an important aspect. Dice's coefficient certainly does that in highlighting the importance of similar attributes over distinguishing ones. In her PhD dissertation, Rodriguez, recognizes that fact in proposing her own *Matching-Distance Model* of similarity between spatial entity classes:

$$S_{MD}(A,B) = \omega_p \cdot S_p(A,B) + \omega_f \cdot S_f(A,B) + \omega_a \cdot S_a(A,B) \tag{8}$$

where ω_p, ω_f and ω_a are weights of the similarity values for parts, functions and attributes respectively of geospatial concepts[15]. Rodriguez uses Tversky's index, tailored to geo-concepts.

3.2.2 Applicability of Feature-based Measures to CBR

Feature-based measures present a solution set that does not need to rely on Geometry to compute similarity. Similarity in this model is established solely on the number of shared and distinctive features of the objects (cases) under consideration. However, the model in its general formulation is insensitive to feature weighting (or salience) in that it only needs feature counts. This consideration must be accounted for by means of appropriate attribute weights and is domain-dependant. For example, apple and bananas may be judged to share the features of being sweet, of being fruits, and opposed in features of color and shape. However, the feature of nutritiousness may be judged to be of less importance than the other features and this would evidently affect the computed similarity.

3.3 Structure-based measures

Case-based reasoning relies on Knowledge Representation Mechanisms (KRMs) to represent the cases. Exemplar KRMs are Description Logics, the Entity-Relational model, Modal logic, etc. Many of these depend on the fundamental relationship of *subsumption* (the `is-a` relationship). Within a case representation, concepts related by means of the subsumption relationship form graphs, revealing the inherent cases structure. Structure-based measures are determined from this apparent structure. It is hypothesized that two concepts present a high degree of similarity if they show a high similarity in their respective graph topologies. A common parent node often serves as a reference point to establish the distance, in edge or node counts, between the concepts and thus their similarity.

We say that

$$sim(C_{current}, C_{CB}) \models f(N_{C_{current}}, N_{C_{CB}}) \rightarrow \mathbb{R} \tag{9}$$

where $N_{C_{current}}$ and $N_{C_{CB}}$ represent nodes or edges count that separate concepts from a certain reference point in the respective structures, often the root node or the least upper bound (closest common parent node).

The model can also be extended to consider similarity between subgraphs, suggesting that more complex cases can be compared for similarity. However, it must be noted that structure-based measures are highly dependant on cases representations or topology. The similarity measure can be greatly affected if different KRMs are used for the case base and the current case representations. Even with a single KRM, there can be significant modeling differences in cases representation that will blur the resulting similarity measure.

3.4 Transformation-based Measures

Another approach to measure similarity between structures (e.g. character strings, gene sequences, images, etc.) is to consider how many operations are needed to *transform* the first structure into the other one. Hamming and Levenshtein distances are classical measures in this realm, labeled *edit distance* measures. The Hamming distance returns the number of symbols that are different between two sequences of equal length. The Levenshtein distance yields the minimum number of edit operations (delete, insert and substitute specifically) needed to morph a sequence into the other one. Many other measures have been derived from Levenshtein's by allowing different sets of edit operations.

Transformation-based measures are particularly useful for comparing structures that refer to the same domain (e.g. DNA sequences). However, they become useless when the structures represent distinct non-overlapping domains. Let's consider for example the strings "`HHTHTTHT`" and "`THTHTTHT`". Hamming's and Levenshtein's distance would both yield a value of 1, suggesting a rather strong similarity considering their length. This makes sense if those strings represent two sequences of 8-tossing of a coin, where "`H`" is "heads" and "`T`" is "tails". However, if those strings represent *codewords* (granted that it is a weird choice for codewords) referring to different concepts, such as *Ship* and *Tank*, then the similarity value of 1 no longer makes sense. Transformation-based measures do not take into consideration the semantic content of the structures. One has to be careful to ensure that the application domain where these measures are to be used is well defined and suffers minimal heterogeneity.

3.5 Information Content-Based Measures

In defining what would become Information Theory, Shannon[16] defined the concept of *information content*, expressed in Equation 10. Information content expresses the degree of likeliness of a specific message coming from an information source. It is therefore intimately linked to the probability of occurrence of that message. If we average the information content over all the K messages generated by the source, we obtain its *entropy* (Equation 11).

$$I_X(x_k) = -\log_b p(x_k) \tag{10}$$

$$H(X) = -\sum_{k=1}^{K} \log_b p(x_k) \tag{11}$$

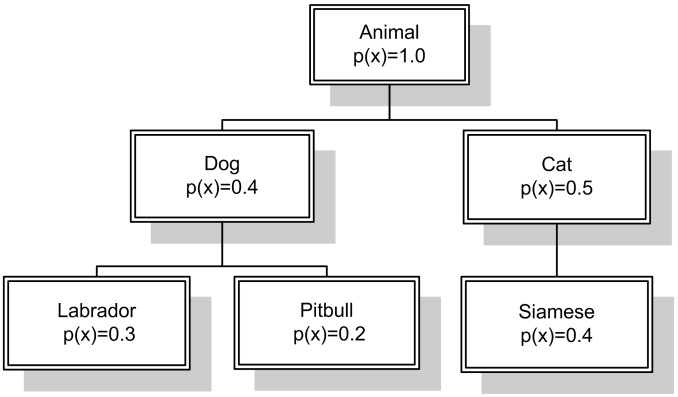

Figure 2: A Taxonomy

To assess the similarity between concepts that belong to the same taxonomy, Resnik proposed to use information content[17]. He reasoned that if we consider the taxonomy to be a source of information, that we could effectively attribute probabilities of occurrence to the concepts of the taxonomy. As such, the root concept's probability value is 1.0 in the sense that any message coming will, at the very least, express the semantic content of that upper concept. It is also natural in his view to consider that the more concrete concepts in the taxonomy are less likely to occur, and thus have lower probabilities. As an example, let's consider Figure 2. Since it is a taxonomy, the instantiation of any concept necessarily involves the root concept of *Animal*, explaining its probability of 1.0. Other probabilities were determined by the specific probabilistic profile of that source[2]. From this, Resnik proposed that the degree of similarity between concepts within a taxonomy corresponds to the information content value of its closest common parent. In Figure 2, the similarity between a *Siamese cat* and a *Dog* is $-\log_2(1.0) = 0$, and the similarity between a *Pitbull* and a *Labrador* is $-\log_2(0.4) = 0.92$ (not normalized).

Resnik's proposal does not take into account all of the tools provided by Information Theory. It opens however the door to a whole new class of similarity measures that could bring light to other important aspects in the establishment of similarity. The key still relies though on proper modeling of the random variables that represent the information source, the *virtual channel* by which a case is semantically mapped to the other (what is the remaining average uncertainty on that mapping?) and their coupling to form a virtual communication system where information loss is to be minimized.

[2]Resnik did not posit that the probabilities of the nodes at the same level should sum to the probability of their closest parent.

4 CBR Measure of Similarity for Command and Control

All of the similarity measure we presented in the previous section have advantages and drawbacks, strengths and weaknesses. Making a choice on a specific approach depends on a number of aspects, the first of which is the actual process that will make use of it. Since CBR was in focus in this work, we ought to bring back the argument that the choice of a similarity is a recognized challenge of this reasoning method (see Section 2.2). Although this challenge could be tackled from the unique viewpoint of CBR research, it is reasonable to think that the application domain for which this reasoning method was chosen may provide insights on what would be the most suitable similarity measure. Command and Control (C2) in Multinational Civil-Military Operations is our application domain. A characterization of its facets, however incomplete, will yield important clues on the applicability of particular measures of similarity. Although such a characterization is out of scope of this paper, we consider that C2 is complex in nature, which would necessarily be reflected in a formal representation. It ensues that the CBR cases will be complex aggregates of atomic concepts. The establishment of cases similarity must thus take place between those aggregates. In this sense, the measures comparing single concepts together either have to grow to consider concept aggregates or be discarded altogether. Geometry-based, feature-based and information content-based classes of measures seem to feature such a growth potential while transformation-based and structure-based seem to be more rigid. This rigidity is explained by the fact that transformation-based measures rely on fixed sets of edit operators that operate on *sequences* or strings; Structure-based measures are highly dependant on the concept graph topology. Since we know that CBR formal cases representation may fluctuate, especially under the C2 application domain, these measures seem not to be the best candidates.

The potential of Geometry-based measures lies in the expression of Equation 2. There is a formal recognition that *concepts* must be ported into a geometric space. Although not formally recognized, those concepts can be aggregates. It is therefore equally possible to compare *ships* with *tanks* (atomic) and *C2 Courses of actions* with one another (complex or molecular). This aspect is also shared by feature-based models where feature extraction is the cornerstone. While this shows the flexibility of such similarity models, it also constitutes their own Achilles' heel in that feature selection is a delicate and critical process.

Resnik's information content-based measure of similarity seems to suffer the same limitation of transformation-based and structure-based approaches because it is set up with an apparent dependance on the topology of the structure. However, Information Theory in its whole brings many other tools that ought to be tested in their proper context. First and foremost, Information Theory reflects on the *uncertainty* inherent to any communication process. Thinking in terms of uncertainty in thus central to the application of this theory. In these terms, we could see how the establishment of similarity is a complex cognitive process that aims at determining how uncertain it is to state that two *things* (let it be objects, CBR cases or C2 Coalition Situations) are identical. This last approach constitute prospective work that is currently being

researched in DRDC Valcartier.

5 Conclusion

Achieving adequate C2 in a contextual environment as heterogeneous and as demanding as Multinational Civil-Military Operations requires appropriate C2 reasoning tools. One such tool, CBR, is actively being researched in DRDC Valcartier. One critical aspect of CBR is the ability to establish similarity between new situations and those expressed in the case base.

This paper exposed many different approaches to measure similarity and discussed their applicability to CBR in the context of complex C2. Further research is being conducted to assess the applicability of Information Theory to this domain.

References

[1] H. Irandoust and A. Benaskeur. *Political, Cultural and Command & Control Challenges in Coalitions*. Canadian Defence Academy Press, In Press 2011.

[2] A. Aamodt and E. Plaza. Case-based reasoning: foundational issues, methodological variations, and system approaches. *Artificial Intelligence Communications*, vol. 7, no. 1:29–59, 1994.

[3] Alexandre Bergeron Guyard. Case-based reasoning for maritime anomaly detection. In *COGnitive systems with Interactive Sensors (COGIS)*, 2010.

[4] Padraig Cunningham. A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21(No. 11):1532–1543, November 2009.

[5] Eric Dorion and Alexandre Bergeron Guyard. Situation analysis for the tactical army commander: Final report. *DRDC Valcartier TR 2010-174*, 2010.

[6] Jean Roy. A knowledge-centric view of situation analysis. *DRDC Valcartier TR 2005-419*, 2007.

[7] Alexandre Bergeron Guyard and Jean Roy. Towards case-based reasoning for maritime anomaly detection: A positioning paper. In *Proceedings of The IASTED International Conference on Intelligent Systems and Control*, 2009.

[8] C.K. Reisbeck and R. Schank. *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, 1989.

[9] L.E. Mujica, J. Vehi, and P. Kolakowski. A hybrid approach of knowledge-based reasoning for structural assessment. *Smart Material Struct.*, 14:1554–1562, 2005.

[10] Robert L. Goldstone and Ji Yun Son. *Similarity*, pages 13–36. Cambridge University Press, 2005.

[11] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In Catriel Beeri and Peter Buneman, editors, *Database Theory – ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 217–235. Springer Berlin / Heidelberg, 1999.

[12] Dominic Widdows. *Geometry and Meaning*. CSLI Publications, 2004.

[13] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2004.

[14] Amos Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[15] Maria Andrea Rodriguez. *Assessing Semantic Similarity among Spatial Entity Classes*. PhD thesis, University of Maine, 2000.

[16] Claude Elwood Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.

[17] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.