A COMPARISON OF SEMANTIC SIMILARITY MODELS IN EVALUATING CONCEPT SIMILARITY

Q.X. Xu*, W.Z. Shi

Dept. of LSGI, The Hong Hong Polytechnic University, Hung Hom, Kowloon, Hong Kong – xu.qx@connect.polyu.hk

Commission II, ICWG II/IV

KEY WORDS: Semantic Similarity, Concept Similarity, Geometric Model, Feature Model, Network Model, Transformational Model

ABSTRACT: The semantic similarities are important in concept definition, recognition, categorization, interpretation, and integration. Many semantic similarity models have been established to evaluate semantic similarities of objects or/and concepts. To find out the suitability and performance of different models in evaluating concept similarities, we make a comparison of four main types of models in this paper: the geometric model, the feature model, the network model, and the transformational model. Fundamental principles and main characteristics of these models are introduced and compared firstly. Land use and land cover concepts of NLCD92 are employed as examples in the case study. The results demonstrate that correlations between these models are very high for a possible reason that all these models are designed to simulate the similarity judgement of human mind.

1. INTRODUCTION

The evaluation of semantic similarities plays an important role in different contexts including: classification definition (Sokal, 1974), categorization (Goldstone and Son, 2005), interpretation (Janowicz, 2008), information retrieval (Janowicz, 2011), and information integration (Hakimpour and Geppert, 2001). Previous studies are mainly conducted by psychologists (Gentner and Markman,1994; Goldstone and Son, 2005), until recently, semantic similarities are highly concerned in Geographic Science for reasons such as requirements of interoperation between different systems (Sheth, 1999).

A number of semantic similarity models have been designed and implemented in consideration of properties, relations or both. Some of them can be applied to evaluate similarities of both concepts and objects, while others can only calculate similarities of objects. A concept is abstracted from a group of individuals with some common characteristics. It is essential for memory and inference to human beings. In order to find out the suitability and performance of different models in computing concept similarity, a comparative analysis is needed.

This paper investigates four main similarity models and gives a comparison of them in evaluating concept similarity. In Section 2, we introduce the basic principles and present at least one classical method for each model. In Section 3, we analyse the main characteristics of these similarity models. Taking three category concepts of NLCD92 classification system as examples, we compare the results of semantic similarity calculated from different models and give a short discussion on the results in Section 4. Finally, Section 5 comes to a conclusion.

2. A BRIEF REVIEW OF SEMANTIC SIMILARITY MODELS

There are mainly five types of semantic similarity models named geometric model, feature model, network model, alignment model, and transformational model. As the alignment model limited by the constraint of one-to-one mapping, which is difficult to apply to evaluate the concept similarity (Schwering, 2008), the other four models will be introduced in this section.

2.1 Geometric Model

A concept in the geometric model is represented as a region in a multidimensional space. Each dimension is a property of the concept, and the range of the dimension represents all possible values of the property. Instead of measuring semantic similarity directly, geometric models measure the semantic distance between concepts. In analogy to spatial distance, a generic formula for the semantic distance measurement is Minkowski Metric (equation 1).

$$d\left(c_1, c_2\right) = \left[\sum_{i=1}^{n} \left|C_{1i} - C_{2i}\right|^r\right]^{1/r} \tag{1}$$

Where c_1 and c_2 are two concepts, n is the number of dimensions used to describe the concept, $C_{1i}(C_{2i})$ is the property value of concept $c_1(c_2)$ in the ith dimension. $r = 2$ results in the Euclidian distance, while $r = 1$ results in the city-block distance.

The similarity is a linear or exponentially decaying function of the distance (Melara et al. 1992). Equation (2) is a possible exponentially decaying function to transform semantic distance to semantic similarity, which results in a similarity normalized between 0 and 1.

$$s\left(c_1, c_2\right) = \frac{1}{1 + d(c_2, c_2)} \tag{2}$$

* Corresponding author. Email: xu.qx@connect.polyu.hk.

2.2 Feature Model

The basis of the feature model is set theory. Property values of a concept are represented as elements in a feature set. Semantic similarity is computed by taking into account both common and distinct features. Common features increase the similarity, while distinct features decrease the similarity. The most famous feature models are Tversky's (1977) Contrast Model (equation 3) and Ratio Model (equation 4).

$$s\left(c_1, c_2\right) = \theta * f(C_1 \cap C_2) + \alpha * f(C_1 - C_2) + \beta * f(C_2 - C_1) \quad (3)$$

$$s\left(c_1, c_2\right) = \frac{f(C_1 \cap C_2)}{f(C_1 \cap C_2) + \alpha * f(C_1 - C_2) + \beta * f(C_2 - C_1)} \quad (4)$$

Where $f()$ is the function either reflecting the salience or prominence of a set of features (Pirró and Euzenat, 2010) or simply determining the cardinality of the set (Schwering, 2006). $C_1 \cap C_2$ is the set of common features of two concepts, while $C_1 - C_2$ ($C_2 - C_1$) is the set of distinct features that belong to concept C_1 (C_2) but do not belong to concept C_2 (C_1). θ, α and β indicate the importance of different components in the similarity estimation. The sum of α and β should equal to 1.

For the contrast model, the similarity value is not bounded between 0 and 1, which makes it difficult for interpretation.

2.3 Network Model

The basis of the network model is graph theory. Concepts are connected through appropriate relations, such as is-a relation, in the semantic network. Concepts are represented by nodes, while the relations between them are represented by edges. The similarity in the network model is calculated by graph-theoretic algorithms. A simple algorithm for semantic distance may be the shortest path model (equation 5).

$$d\left(c_1, c_2\right) = \min length(P_{c_1,c_2}) \quad (5)$$

Where P_{c_1,c_2} is the length between C_1 and C_2.

For a semantic network with only is-a relations, similarity values in the network model is highly sensitive to the predefined hierarchy network (Rodríguez, 2000).

2.4 Transformational Model

The similarity in the transformation model is equal to the number of transformations to make one concept identical to the other concept (Hahn et al. 2009). When needed transformations increase, the similarity decreases monotonically. Transformational operations may be counted based on a coding language (Hodgetts et al. 2009). According to Kolmogorov complexity theory, the number of transformations is the smallest number of operations that the computer program transforms one concept into the other (Goldstone, 1999).

Indeed, the calculation of the semantic similarity from the transformational model needs some comparisons (Grimm et al. 2012). Transformational operations are only conducted on distinct components, while identical components are not considered.

3. CHARACTERISTICS OF SEMANTIC SIMILARITY MODELS

3.1 Similarity and Dissimilarity

The semantic similarity is obtained directly or converted and normalized from the semantic dissimilarity (distance) indirectly.

The feature model is a typical representative of the former category. Common and distinct features are combined to evaluate semantic similarities. The more common features and the less distinct features, the higher is the overall semantic similarity.

Different kinds of distance, including spatial distance, path length, and transformational complexity, are considered in the geometric model, the network model and the transformational model. There is a negative relationship between similarity and distance. The shorter the distance, the higher the similarity.

3.2 Properties and Relations

Properties describe the characteristics of a concept, while relations describe connections between concepts.

Properties are considered in all models except for the pure network model. The name and the range of the property are explicitly represented in the geometric model. Property values are arranged along the dimension in some rank. In the feature model, property values are simply listed in the feature set. In the transformational model, properties are aligned into two types of properties: matched property and unmatched property. Normally, less number of operations is needed to transform between the matched properties than the unmatched.

Relations are considered in the network model and the transformational model. Relations in the network model are usually hierarchic or associative, while relations are aligned for transformation in the transformational model.

3.3 Metric and Non-metric

The metric character is the most important assumptions of the geometric model. In metric space, the semantic distance meets the three metric axioms (Gärdenfors, 2000): minimality (equation 6), symmetry (equation 7), and triangle inequality (equation 8).

$$d\left(c_1, c_2\right) = 0 \Rightarrow c_1 = c_2 \wedge d\left(c_1, c_2\right) \geq 0 \quad (6)$$

$$d\left(c_1, c_2\right) = d\left(c_2, c_1\right) \quad (7)$$

$$d\left(c_1, c_2\right) + d\left(c_2, c_3\right) \geq d\left(c_1, c_3\right) \quad (8)$$

The axiom of minimality indicates that if the distance between two concepts equals 0, then the concepts are identical. The axiom of symmetry indicates that the order of concepts does not affect the magnitude of distance. The axiom of triangle inequality indicates that the direct distance from one concept to the other is not larger than the sum of the distance from any one of them to an intermediate concept.

The geometric model is set up on the metric space, which has been criticized for disagreement with human cognitive process. Hence, the feature model is designed which discards the metric character. Network models hold the axiom of minimality and triangle inequality, while undirected network models keep symmetric and directed network models are asymmetric. For the transformational model, it holds the metric character except for the symmetry.

The metric characteristics of the mentioned models are expressed in Table 1. It is clear that the main discrepancy is concentrated on the symmetric axiom, which is preserved by two models and disagreed by the other three.

	SYMMETRIC	MINIMALITY	TRIANGLE INEQUALITY
GEOMETRIC MODEL	Yese	Yes	Yes
FEATURE MODEL	No	No	No
NETWORK MODEL (UNDIRECTED)	Yes	Yes	Yes
NETWORK MODEL (DIRECTED)	No	Yes	Yes
TRANSFORMATIONAL MODEL	No	Yes	Yes

Table 1 Metric Characteristics of Semantic Similarity Models

3.4 Degree of Similarity

The degree of similarity refers to details of information employed to calculate semantic similarities. In other words, it indicates the ability to separate two concepts.

Properties of a concept in the geometric model are explicitly represented as dimensions. The range of a dimension includes all possible values of the property, which may be nominal values, interval values, ordinal values, or ratio values. Different property values of the concepts within a dimension can have influences on the overall similarity.

For the feature model, the property values are represented as elements in the set. Property information may be abandoned or implicitly presented as compound elements. For example, in Section 4.3, "Deciduous Tree" is a property value from which we may not be able to infer its original property name, while "Dominated area: [75,100]" describe its property name "Dominated area" implicitly with a compound element. From a feature model, the range of properties cannot be acquired either.

Instead of considering properties, only relations between concepts are employed to evaluate similarity in a pure network model. However, a problem is that how the relations can be acquired if no information on relations is included in a concept. In authors' opinion, properties need to be extracted and act as a basis to establish concept relations.

The transformational model considers both properties and relations which are mainly perceivable, because we need to transform from one to the other. For example, the transformation may occur from the property of "Deciduous Tree" to "Evergreen Tree" or from the relation of "Horizontal" to "Vertical".

4. CASE STUDY: SEMANTIC SIMILARITY OF LAND USE AND LAND COVER CONCEPT

4.1 Experiment Data

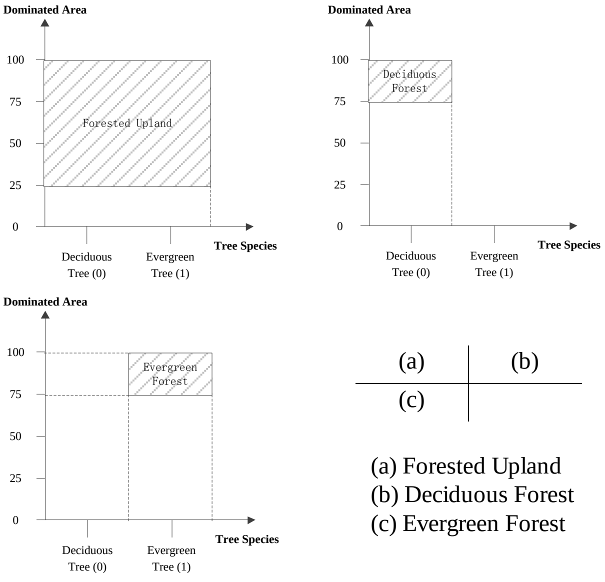
As running examples we use three land use and land cover concepts of National Land Cover Data 1992 classification system (NLCD92), which is an II Level classification system modified from the Anderson Land use and Land Cover Classification System (EPA, 2008). The definitions of them are as follows.

> *Forested Upland - Areas characterized by tree cover (natural or semi-natural woody vegetation, generally greater than 6 meters tall); tree canopy accounts for 25-100 percent of the cover.*
>
> *Deciduous Forest - Areas dominated by trees where 75 percent or more of the tree species shed foliage simultaneously in response to seasonal change.*
>
> *Evergreen Forest - Areas dominated by trees where 75 percent or more of the tree species maintain their leaves all year. Canopy is never without green foliage.*

"Forested Upland" is a First-Level concept, while "Deciduous Forest" and "Evergreen Forest" are sub-concepts of "Forested Upland" in the Second-Level of NLCD92.

4.2 Experimental Results: The Geometric Model

In geometric models, running examples can be represented as Figure 1. Concepts are mapped into a two-dimensional space: one dimension is "Tree Species" and the other is "Dominated Area". The dimension of Tree Species is a nominal dimension with the range of {Deciduous Tree, Evergreen Tree}, while the dimension of Dominated Area is a ratio dimension with the range of [0, 100%]. The range of the super-concept covers its sub-concepts in this situation, which can be demonstrated from a comparison of dashed areas in Figure 1.



(a) Forested Upland
(b) Deciduous Forest
(c) Evergreen Forest

Figure 1 Representation of Running Examples in the Geometric Model

As a concept is a region in the space, semantic distance can be equal to the distance between the region centroids. Results of the semantic distance and the semantic similarity of running examples are expressed in Table 2.

	Forested Upland	Deciduous Forest	Evergreen Forest
Forested Upland	0.00 (1.00)	0.35 (0.74)	0.35 (0.74)
Deciduous Forest	0.35 (0.74)	0.00 (1.00)	0.50 (0.67)
Evergreen Forest	0.35 (0.74)	0.50 (0.67)	0.00 (1.00)

Table 2 Semantic Distance (Semantic Similarity) in the Geometric Model

The self-similarity, equal to 1, is the maximum along the diagonal. The similarity between sub-concepts (0.67) is smaller than that between the super-concept and a sub-concept (0.74) in the running examples. However, an opposite result can occur, especially in the situation when many sub-concepts inherit from one super-concept.

4.3 Experimental Results: The Feature Model

In the feature model, running examples are represented as unstructured feature sets, as illustrated in Figure 2. "Forested Upland" possesses four elements, while "Deciduous Forest" and "Evergreen Forest" possess two respectively.

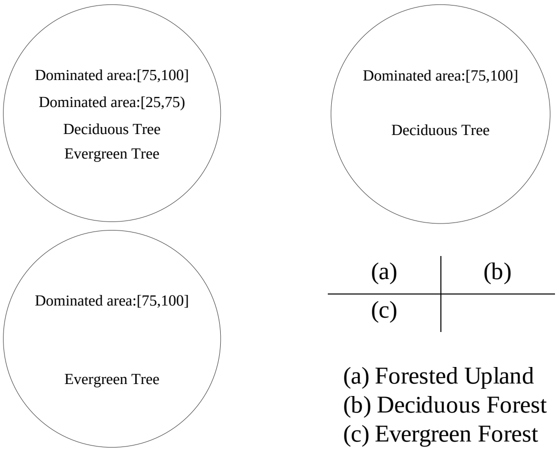


Figure 2 Representation of Running Examples in the Feature Model

Properties, which are employed to characterize the concept, are not, or at least not explicitly, represented in the feature model, unless compound feature elements are used, e.g. "Dominated area: [75,100]".

Results of the semantic similarity of running examples in the contrast model and ratio model are expressed in Table 3, with $\theta = 1$, $\alpha = 0.7$ and $\beta = 0.3$.

	Forested Upland	Deciduous Forest	Evergreen Forest
Forested Upland	4.0 (1.0)	3.4 (0.6)	3.4 (0.6)
Deciduous Forest	2.6 (0.8)	2.0 (1.0)	2.0 (0.5)
Evergreen Forest	2.6 (0.8)	2.0 (0.5)	2.0 (1.0)

Table 3 Semantic Similarity in the Contrast (Ratio) Model

In accordance with the geometric model, similarities of the ratio model are maximal and equal to 1 along the diagonal. For the contrast model, the situation becomes complex. The self-similarity may be not the maximum, for example, the self-similarity of "Deciduous Forest" (2.0) is smaller than that between "Deciduous Forest" and "Evergreen Forest" (2.6). Additionally, the similarity in the contrast model is not normalized to [0, 1], which makes it hard for interpretation.

4.4 Experimental Results: The Network Model

Instead of representing properties, relations between concepts are represented in the network model. In the running examples, only the is-a relation is employed, which leads to a simple undirected semantic network, as illustrated in Figure 3.

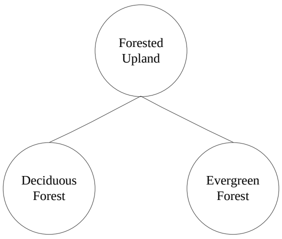


Figure 3 Representation of Running Examples in the Undirected Network Model

If weights of all edges are identical and equal to 1, the results of the shortest path between concepts are presented in Table 4. The semantic similarity is calculated and normalized according to Equation (2).

	Forested Upland	Deciduous Forest	Evergreen Forest
Forested Upland	0 (1.0)	1 (0.5)	1 (0.5)
Deciduous Forest	1 (0. 5)	0 (1.0)	2 (0. 3)
Evergreen Forest	1 (0. 5)	2 (0. 3)	0 (1.0)

Table 4 Semantic Distance (Semantic Similarity) in the Undirected Network Model

The self-similarity value is still the maximum and equal to 1 in the network model. The similarity between sub-concept nodes of the same super-concept node (0.3) is smaller than that between the node and its super-concept node (0.5).

In a hierarchic structure, general concepts (e.g. "Forested Upland") are located in higher hierarchic levels, while detailed concepts (e.g. "Deciduous Forest") are located in lower hierarchic levels. The similarity of concepts with a common super-concept in a closer hierarchic level should be larger than that with a common super-concept far away from hierarchic levels of themselves.

4.5 Experimental Results: The Transformational Model

Other than representing properties or relations directly, the transformational model concerns how to transform one concept into the other. A coding language is employed to represent concepts. Property values of concepts are labelled by English letters, here, with "A", "B", "C", and "D". Thereafter, concepts can be represented as a letter list. For example, "ABCD" represents "Forested Upland", "AC" represents "Deciduous Forest", and "AD" represents "Evergreen Forest". We can transform these concepts to each other with three operations: delete, create, and apply, as illustrated in Figure 4. For example, "Forested Upland" can be transformed into "Deciduous Forest" by the operation of "delete" two times: delete the property of "B" and delete the property of "D".

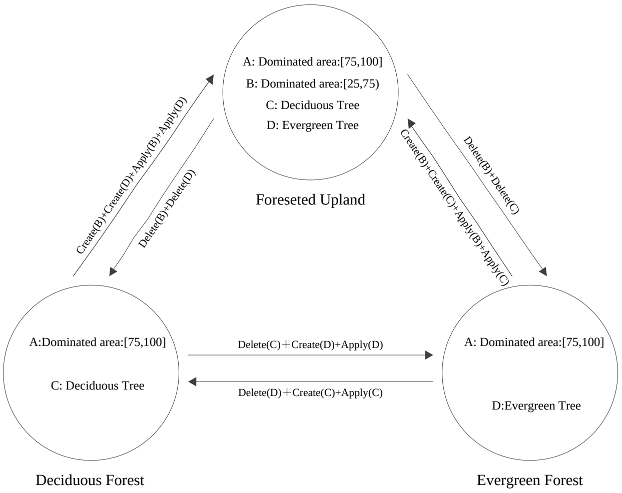

A: Dominated area:[75,100]
B: Dominated area:[25,75)
C: Deciduous Tree
D: Evergreen Tree
Foreseted Upland
Create(B)+Create(D)+Apply(B)+Apply(D)
Delete(B)+Delete(D)
Delete(B)+Delete(C)
Create(B)+Create(C)+Apply(B)+Apply(C)
A:Dominated area:[75,100]
C: Deciduous Tree
Deciduous Forest
Delete(C)＋Create(D)+Apply(D)
Delete(D)＋Create(C)+Apply(C)
A: Dominated area:[75,100]
D:Evergreen Tree
Evergreen Forest

Figure 4 Representation of Running Examples in the Transformational Model

The arrow-line indicates the direction of the transformation. It can be seen that the number of operations is not symmetric. Usually, the transformation from a general concept to a detailed concept needs fewer operations than versus.

The results of the semantic distance and the normalized semantic similarity are expressed in Table 5. The self-similarity is maximal and equal to 1 along the diagonal as well.

	Forested Upland	Deciduous Forest	Evergreen Forest
Forested Upland	0 (1.00)	2 (0.33)	2 (0.33)
Deciduous Forest	4 (0.20)	0 (1.00)	3 (0.25)
Evergreen Forest	4 (0.20)	3 (0.25)	0 (1.00)

Table 5 Semantic Distance (Semantic Similarity) in the Transformational Model

4.6 Comparison of Experimental Results

To present semantic similarity values of different models visually, the results are showed as Figure 5. The horizontal axis represents concept-pairs with the initial of each concept. For example, "FD" represents the semantic similarity between "Forested Upland" and "Deciduous Forest". Since that similarity values of the contrast model are not normalized to [0, 1], only results of the other four models are represented.

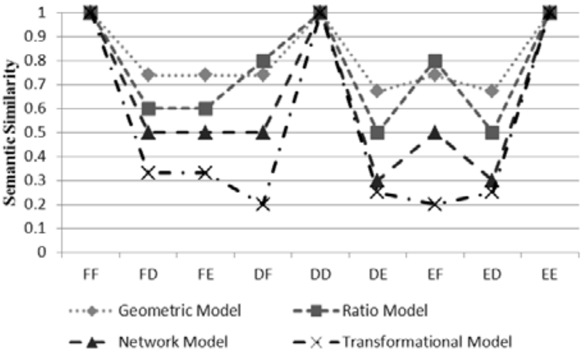


Figure 5 Semantic Similarity of Running Examples in different Models

The diagram demonstrated that the values of self-similarity in all four models are maximal to 1. For the running examples, the minimum value occurs at "DE"/"ED", "DE"/"ED", "DE"/"ED", and "DF"/"EF" for the geometric model, the feature model, the network model, and the transformational model respectively. On average, the magnitude of the similarity values descends in the sequence from the geometric model, the feature model, the network model, to the transformational model.

From the diagram, it seems that the trends of all four models are similar. To validate this, we employ Pearson's r to reflect the strength of linear relations between them (Table 6).

	Geometric	Ratio	Network	Transformational
Geometric	1.00	0.92	1.00	0.98
Ratio	0.92	1.00	0.93	0.82
Network	1.00	0.93	1.00	0.96
Transformational	0.98	0.82	0.96	1.00

Table 6 Correlation of Running Example Semantic Similarity in different Models

From Table 6, it is clear that correlations are all very high with a minimum value of 0.82 between the ratio model and the transformational model. It demonstrates that there is a strong positive linear relationship between these models. This should be so, as all models are designed to simulate the same process, the concept similarity in human mind. Although we can induce that there exist high correlations between models, a value of 1 may not indicate the relation is linear because only three concepts are applied in the running examples.

5. CONCLUSIONS

This paper gives a comparison of four different types of semantic similarity models in concept similarities. In general, there is no unique representation for a concept. The feature model combines both common and distinct features to evaluate similarity directly, while other three models employ distance to separate similarity indirectly. The spatial distance, path length, and transformational complexity are measured in the geometric model, the network model and the transformational model respectively. The geometric model and the feature model only take properties of concepts into account, while the pure network model only take relations into account. Both properties and relations can be applied in the transformational model. The geometric model is a metric model, while the feature model is nonmetric. The network model and the transformational model are asymmetric, but hold metric axiom of minimality and triangle inequality.

From the running examples of three land use and land cover concepts of NLCD92, the results show that similarities calculated from these models (except for the contrast model) accord with each other very well. Based on the analysis, it is reasonable that all models are designed to simulate an identical process of the concept similarity assessment in human mind.

One possible trend of future researches on models for concept similarity evaluation is to integrate merits of different models into a hybrid model. Some researchers have made some contributions on this point. For example, Schwering (2005) propose a hybrid model to integrate merits of the geometric model and the network model.

ACKNOWLEDGEMENT

The work presented in this paper is supported by Ministry of Science and Technology of China (Project No.: 2012BAJ15B04, 2012AA12A305) and The Hong Kong Polytechnic University (Project No.: 1-ZV4F, G-U753, G-YF24, G-YJ75, G-YG66).

REFERENCES

EPA, (2008). AERSURFACE User’s Guide. EPA-454/B-08-001.

Gärdenfors, P. (2000). Conceptual Space: The Geometry of Thought. pp. 17-18. Cambridge, MA: MIT Press.

Gentner, D. and Markman, A.B. (1994). Structural Alignment in Comparison: No Difference without Similarity. Psychological Science, 5: 152–8.

Goldstone, R.L. (1999). Similarity. In: Wilson, R.A. and Keil, F.C. (eds.), Mit Encyclopedia of the Cognitive Sciences. pp. 763-765. Cambridge, MA: MIT Press.

Goldstone, R. L. and Son, J. (2005). Similarity. In: Holyoak, K. and Morrison, R. (eds.), Cambridge Handbook of Thinking and Reasoning, pp. 13–36. Cambridge University Press. doi:10.2277/0521531012.

Grimm, L.R., Rein, J.R. and Markman, A.B. (2012). Determining Transformation Distance in Similarity: Considerations for Assessing Representational Changes A Priori. Thinking & Reasoning, 18(1): 59-80.

Hahn, U., Close, J. and Graf, M. (2009). Transformation Direction Influences Shape Similarity Judgements. Psychological Science, 20: 447-454.

Hakimpour, F. and Geppert, A. (2001). Resolving Semantic Heterogeneity in Schema Integration: An Ontology Based Approach. In: Welty, C. and Smith, B. (eds.), Proceedings of International Conference on Formal Ontologies in Information Systems FOIS'01. ACM Press.

Hodgetts, C.J., Hahn, U. and Chater, N. (2009). Transformation and Alignment in Similarity. Cognition, 113: 62-79.

Melara, R.D., Marks, L.E. and Lesko, K.E. (1992). Optional Processes in Similarity Judgements. Perception & Psychophysics, 51(2): 123-133.

Janowicz, K. (2008). Kinds of Contexts and Their Impact on Semantic Similarity Measurement. In Proc. 5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea), 6th IEEE International Conference on Pervasive Computing and Communication (PerCom), IEEE Computer Society. doi:10.1109/PERCOM.2008.35.

Janowicz, K., Raubal, M. and Kuhn, W. (2011). The Semantics of Similarity in Geographical Information Retrieval. Journal of Spatial Information Science, 2:29-57.

Pirró, G. and Euzenat, J. (2010). A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In: Patel-Schneider et al. (eds.), ISWC 2010, Part I. LNCS, vol. 6496, pp. 615-630. Springer, Heidelberg (2010).

Rodríguez, M.A. (2000). Assessing Semantic Similarity among Spatial Entity Classes. Ph.D. Thesis, University of Maine.

Schwering, A. (2005). Hybrid Model for Semantic Similarity Measurement. In Proceedings of the Fourth International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE05), Agia Napa, Cyprus.

Schwering, A. (2006). Semantic Similarity Measurement including Spatial Relations for Semantic Information Retrieval of Geo-Spatial Data. Ph.D. Thesis, University of Münster.

Schwering, A. (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. Transactions in GIS, 12(1): 5-29.

Sheth, A. (1999). Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. In: Goodchild, M., Egenhofer, M., Fegeas, R. and Kottman, C. (eds.). Interoperating Geographic Information Systems. pp. 2-30. Kluwer Academic Publisher, Norwell, MA.

Sokal, R.R. (1974). Classification: Purpose, Principles, Prospects. Science, 185 (4157): 1115-1123

Tversky, A. (1977). Features of Similarity. Psychological Review, 84: 327-352.