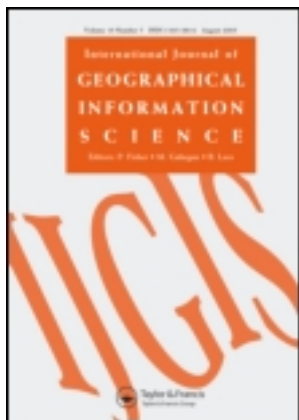


This article was downloaded by: [Arizona State University]

On: 21 March 2013, At: 18:03

Publisher: Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



International Journal of Geographical Information Science

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/tgis20>

Semantic similarity measurement based on knowledge mining: an artificial neural net approach

Wenwen Li ^a, Robert Raskin ^b & Michael F. Goodchild ^a

^a Center for Spatial Studies, University of California, Santa Barbara, CA, 93106, USA

^b Jet Propulsion Laboratory, Pasadena, CA, 91109, USA

Version of record first published: 15 Feb 2012.

To cite this article: Wenwen Li, Robert Raskin & Michael F. Goodchild (2012): Semantic similarity measurement based on knowledge mining: an artificial neural net approach, International Journal of Geographical Information Science, 26:8, 1415-1435

To link to this article: <http://dx.doi.org/10.1080/13658816.2011.635595>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Semantic similarity measurement based on knowledge mining: an artificial neural net approach

Wenwen Li^{a*}, Robert Raskin^b and Michael F. Goodchild^a

^aCenter for Spatial Studies, University of California, Santa Barbara, CA 93106, USA; ^bJet
Propulsion Laboratory, Pasadena, CA 91109, USA

(Received 15 June 2011; final version received 16 October 2011)

This article presents a new approach to automatically measure semantic similarity between spatial objects. It combines a description logic based knowledge base (an ontology) and a multi-layer neural network to simulate the human process of similarity perception. In the knowledge base, spatial concepts are organized hierarchically and are modelled by a set of features that best represent the spatial, temporal and descriptive attributes of the concepts, such as origin, shape and function. Water body ontology is used as a case study. The neural network was designed and human subjects' rankings on similarity of concept pairs were collected for data training, knowledge mining and result validation. The experiment shows that the proposed method achieves good performance in terms of both correlation and mean standard error analysis in measuring the similarity between neural network prediction and human subject ranking. The application of similarity measurement with respect to improving relevancy ranking of a semantic search engine is introduced at the end.

Keywords: semantic similarity; geospatial semantics; geospatial knowledge discovery; spatial data mining; search engine; ranking; ontology

1. Introduction

Semantic similarity is an important notion with two dimensions. It is used to describe the semantic distance between two concepts (either within a single ontology or among two different ontologies) or to measure the semantic distance between a concept and a word or a term. A 'concept' is a cognitive unit of meaning and is always represented with a class within an ontology; a 'word' is a natural language element comprising information in a user query or in a document on the World Wide Web (WWW). Sometimes, people use the terms 'similar' and 'related' interchangeably because both are used to measure 'relatedness'. However, each concept focuses on different aspects of relatedness. For example, a car is more *related* to gasoline than to a bike, whereas a car is more *similar* to a bike than to gasoline. Identifying semantic relationships focuses on *qualitatively* measuring the structural relationships of concepts in an explicit hierarchy, such as parent–child, synonyms. Semantic similarity measures how closely two concepts are related by providing a *quantitative* value. The objective of the article is to apply a 'machine expert' to simulate the

*Corresponding author. Email: wenwen@spatial.ucsb.edu

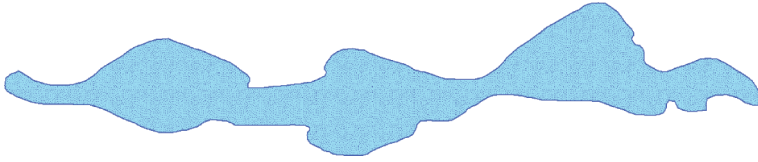


Figure 1. Vagueness in water features: three lakes or a meandering river? (Santos *et al.* 2005).

human perception process in measuring the semantic similarity between spatial objects quantitatively. The hydrology domain will be the context of this study.

Santos (Santos *et al.* 2005) demonstrated an example of conceptual vagueness in water features, which is a ubiquitous problem in geospatial domain. Figure 1 may denote three lakes connected by channels or a river with narrow and broad stretches. The decision from one interpretation to another depends on the existence of: (1) a complete set of information describing the above water feature, (2) a clear semantic definition of the concepts of 'lake' and 'river' to describe the boundaries of applicability of both terms and (3) an effective algorithm that can precisely match the given feature to an existing concept by measuring the similarities based on the information provided in (1) and (2). Practical solutions applying similarity measurement theories to answer these three research questions are the focus of this article.

Similarity measurement theories stem from psychological studies of the human ability to intuitively determine how similar two objects are and to quantify the similarity with a relation (Kebler 2007). In the late 1980s, computer scientists in the field of AI engaged in this research, focusing on building computational models of ambiguous reasoning. With the invention of the Semantic Web (Berners-Lee *et al.* 2001), researchers have attempted to combine similarity research with semantic technologies. Semantic similarity is central to many cognitive processes and plays an important role in how humans process and reason about information (Medin *et al.* 1993, Gentner and Markman 1995, Goldstone and Son 2005, Schwering and Kuhn 2009). Similarity enables semantic interoperability between distributed information systems and web resources, thereby improving the quality of retrieval tasks for Internet users (Cilibrasi and Vitanyi 2006, Janowicz 2006, Janowicz *et al.* 2011). Because of the growth of heterogeneous and independent data repositories, similarity-based information processing has become essential to data/knowledge discovery among distributed data repositories by providing a measure of the degree of relatedness between concepts (Sneth 1999). Hence, the measurement of 'semantic similarity' has emerged as an important topic in several areas of research.

A variety of applications are benefiting from similarity research, as large numbers of practical questions relate to disambiguation and distinction of concepts. For example, Google by 5 October 2011 had received in total 289,000,000 questions asking about the difference between one concept and another and Microsoft Bing search by 5 October 2011 had received in total 154,000,000 such questions. In hydrological science, semantic similarity is often used to identify objects that are conceptually close. According to Santos (Santos *et al.* 2005), an explicit model expressing a complete lattice would greatly help to eliminate the intrinsic vagueness and ambiguity of water features. In geospatial information science, the ontological modelling and similarity identification between geometric characteristics of a single object and geographic relationships between spatial objects help to improve the effectiveness of map generalization. In the web search field, traditional search engines are susceptible to the problems posed by the richness of natural language, especially the

multitude of ways that the same concept can be expressed. Therefore, it is not possible to return satisfying results when making an effort to directly match user query terms with the database. Similarity measurement also provides a way to improve relevance ranking by eliminating conceptual ambiguities existing in user queries and metadata of documents (Resnik 1999, Losif and Potamianos 2007).

Though it is fast and precise for a computer to process the binary equivalence or non-equivalence of two entities, the computation of similarity is a complex and non-trivial problem (Schwering 2008). In the following sections, we review the literature and then discuss the build-up of ontological data, the methodology in use and, finally, the experiments conducted to validate the proposed methodology.

2. Previous work

Existing semantic similarity methods fall into three categories. In general, they can be categorized as edge-counting techniques (Eshera and Fu 1984, Rada *et al.* 1989, Budanitsky and Hirst 2001), information theory based models (Richardson *et al.* 1994, Lin 1998, Resnik 1999, Seco *et al.* 2004) and feature matching models (Tversky and Gati 1978, Sattath and Tversky 1987, Tversky 1977; Rodriguez and Egenhofer 2003, 2004).

Edge-counting techniques are based on a network of semantic relations between concepts and involve calculation of the edge distances between objects in that network. A drawback of edge-counting is its difficulty in defining link distance in a uniform manner. In a practical knowledge base (KB), the distance between terminologies varies dramatically between categories and sub-categories, especially when some categories are much denser (have more subclasses) than others.

Information theory based models measure maximal information shared by two objects, calculated by the negative log likelihood of the shared information. In this measurement, when probability increases, the informativeness decreases. So the higher the level a concept is, the higher the probability is, and thus the lower the information content it has. The statistics-based method lacks semantic support in the similarity measurement and therefore has a bias of human judgement.

In comparison to the above methods, the family of feature-based models, also called classic models, is the most prominent approach for similarity measurement. This approach is object oriented and describes resources by a set of features, such as components (roof and doors) and functionalities (residential or commercial use). The similarity between objects is a function of common and distinguishing features. For example, the Matching Distance Similarity Measure (MDSM) (Rodriguez and Egenhofer 2004) is a feature-based model to measure the similarity between spatial entities. MDSM considers two kinds of features: functional features and descriptive features. The similarity calculation for each feature type counts the common and differential features, and then applies them into Tversky's ratio model. The overall similarity is the linear sum of the weighted similarity values for each feature type.

Although the prominence of a feature is a deterministic factor in human measurements of similarity, current feature-based models, which are still based on a knowledge base with a simple logic, are not suitable for mainstream knowledge representation, such as First Order Logic (FOL) and Description Logic (DL). This is because in the DL, more semantic constraints are used to restrict an object and the interrelations between objects than that in a taxonomy. And the structure of knowledge base changes from a tree-based structure to an interconnected graph. Therefore, a new method is needed to adopt the advancement of knowledge representation (d'Amato *et al.* 2008). The similarity equation in the MDSM

model is a linear product of multiple feature sets. In contrast, human recognition of similarity is sometimes too complex to be simulated by these mathematical equations. We cannot rely on humans to provide the similarity for all the facts in the world because it would be too time-consuming and inflexible. Instead, there is a need for a ‘machine expert’ to simulate the human perception process. This capability requires the machine to have the ability to learn how to carry out tasks based on initial human experience. In the next section, we will discuss in detail the proposed method in applying artificial neural network (ANN) to measure the similarity between spatial objects automatically. The build-up of a comprehensive knowledge base and the training process will be introduced as well.

3. Proposed methodology

Artificial Neural Networks (ANNs), or neural nets, have a remarkable ability to derive patterns from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. Another desirable feature of neural nets is their ability to learn. This capability is undoubtedly the reason that several search engines (e.g. MSN search) utilize the ANN model. With a neural net, an improved set of results could be produced over time to provide more relevant search results. The ANN model is inspired by biological neural networks, in which the human brain’s information-processing capability is thought to emerge from a network of neurons. Since the 1940s, computer scientists have tried to model abstract ‘neurons’ to capture the essence of neural computation in the brain. A neural net is strongly self-organized and can create its own representation of the information received during a learning period. Meanwhile, it is highly tolerant of noisy data and faults (Singh and Chauhan 2005), which is very important to our application as human evaluation may have a big bias as well. Of the family of ANN algorithms, the *Multiple Layer Feed-Forward Neural Network* (MLFFN) is quite popular because of its ability to model complex relationships between output and input data. Adding more hidden units to the network makes it possible for MLFFN to represent any continuous, or even discontinuous, functions of the input parameters. This is an advantage of MLFFN over other statistical algorithms proposed in the literature. In this article, MLFFN is utilized to measure semantic similarity automatically.

3.1. MLFFN algorithm

An MLFFN is used here to conduct numerical learning to simulate the knowledge propagation in a biological neural network. Figure 2 shows a general design of the multi-layer neural net that has multiple inputs and outputs.

The core of the algorithm is back propagation and forward propagation, where back propagation is used to train the neural net to get a stable transition matrix W , which transits information from input nodes to hidden nodes, and V , which transits information from hidden nodes to output nodes. Forward propagation is used to measure the difference between predicted output and the desired output using current W and V . The adopted error metric is the mean squared error (E) between output (O_i) and desired correct output (C_i):

$$E = \frac{1}{n} \sum_{i=1}^n (C_i - O_i)^2. \quad (1)$$

The detailed algorithm is as follows:

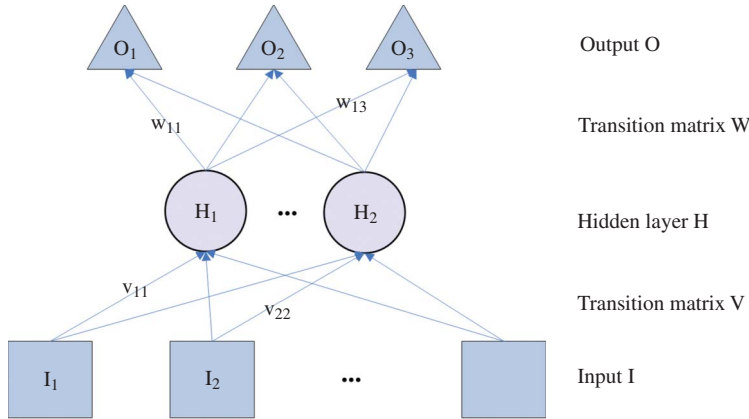


Figure 2. Design of a MLFNN.

- (1) Initialize W and V with given boundaries
- (2) Randomize given data D (a set of input vectors)
- (3) For each element in D ,

a. perform back propagation by

$$\Delta W_{ij} = -\alpha \frac{\partial E}{\partial W_{ij}} = \alpha (C_i - O_i) O_i (1 - O_i) I_j \quad (2)$$

$$\Delta V_{jk} = \sum_i W_{ij} \Delta W_{ij} (1 - H_j) I_k \quad (3)$$

$$W = W + \Delta W \quad (4)$$

$$V = V + \Delta V \quad (5)$$

b. perform forward propagation as follows:

$$H_j = \sigma \left(\sum_k V_{jk} I_k \right) \quad (6)$$

$$O = \sigma \left(\sum_j W_{ij} H_j \right) \quad (7)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$

c. calculate the mean squared error between each output and desired output. If the worst error is lower than a given good-minimum-error, then the network has completed the training, returning V and W as two transition matrices. If the error is not lower than the given good-minimum-error, the algorithm will repeat back propagation to continue training the network.

3.2. *The acquisition of prior knowledge*

Prior knowledge acts as the training dataset for the neural net. It determines how well the transition matrices could be built in the machine learning process. Although a neural net is highly tolerant of noisy data, completeness and representativeness of the prior knowledge is still significant for the accuracy (the closeness of a machine measurement to human measurement) of semantic classification in similarity measurement. A neural net requires that the representation of knowledge be complete because the training process relies on the explicitly defined knowledge. Any uncertainty in the knowledge definition may lead to the failure of the predictive capability of the neural net. Of all the existing machine languages, DL is able to define a complete knowledge concentrating on a specialized application area. It formalizes the knowledge structure by retaining an emphasis on definitions and properties of categories. DL is based on the closed world assumption, by which those ground atomic sentences not asserted to be true are assumed to be false. Therefore, DL is suitable to represent knowledge and to build a domain knowledge base or an ontology. The knowledge structure and content defined in the ontology should be representative and reflect characteristics of an application domain, such as the geospatial domain. As discussed in Section 1, humans tend to measure similarity by comparing features of domain concepts, so to emulate human behaviour a neural net requires a complete feature set that corresponds to the concepts defined in the knowledge base. Sometimes only a few prominent features determine the similarity, rather than the combination effects of all the features. Therefore, besides the completeness in the definition of feature types, the definition of prominent features of the concepts in a certain domain is also important.

In this article, the knowledge base of water bodies together with their spatial, temporal and descriptive features/properties is introduced. The sources of the knowledge base include SWEET ontology (Raskin and Pan 2005), CUAHSI ontology (<http://www.cuahsi.org>), GeoWordNet (<http://geowordnet.semanticmatching.org/>), USGS Water Science Glossary of Terms (<http://ga.water.usgs.gov/edu/dictionary.html>) and Wikipedia (<http://wikipedia.org>). Figure 3 shows all the water body concepts that are modelled ontologically in this research. Concepts such as 'River', 'Creek' and 'Sea' forming the inner circle are the core water body concepts used for machine-based training and learning of similarity. The peripheral concepts, such as 'Burn' and 'Draw', are modelled but not yet considered in measuring the similarity with other objects in the current phase because of the limited information encoded about them in the sources of water body ontologies/articles.

Figure 4 shows the ontological framework of water body described using 17 features; the meaning of each is listed in Table 1. The notations on the arrows connecting the 'WaterBody' node and all the green nodes in Figure 4 are the features that describe a water body object. Other nodes that are descendants of the green nodes are the objects that a 'WaterBody' has with a certain predicate, for example a water body may have functions {Irrigation, PublicSupply, Recreation, PassengerExchange, EcologicalFlow, PowerPlant, Industry, Building/RepairingBoats, Wildlife, Aquaculture, FloodProtection, Mining, LiveStock, ShipShelter, HydroElectricPower, TransferringCargo, WaterQualityImprovement, ShorelineErosion and Aesthetics}. Each member in this set can be used to replace 'Function' in the triple expression {'WaterBody', 'hasFunctionality', Function}. For example, a triple expression {'WaterBody', 'hasFunctionality', 'PublicSupply'} means that a type of 'WaterBody' can be used for supplying water to the public. Any object that is a subclass of 'WaterBody' can be applied to the framework with the specific attribute value defined for that object.



Downloaded by [Arizona State University] at 18:03 21 March 2013

Downloaded by [Arizona State University] at 18:03 21 March 2013

Downloaded by [Arizona State University] at 18:03 21 March 2013

Downloaded by [Arizona State University] at 18:03 21 March 2013

- Downloaded by [Arizona State University] at 18:03 21 March 2013

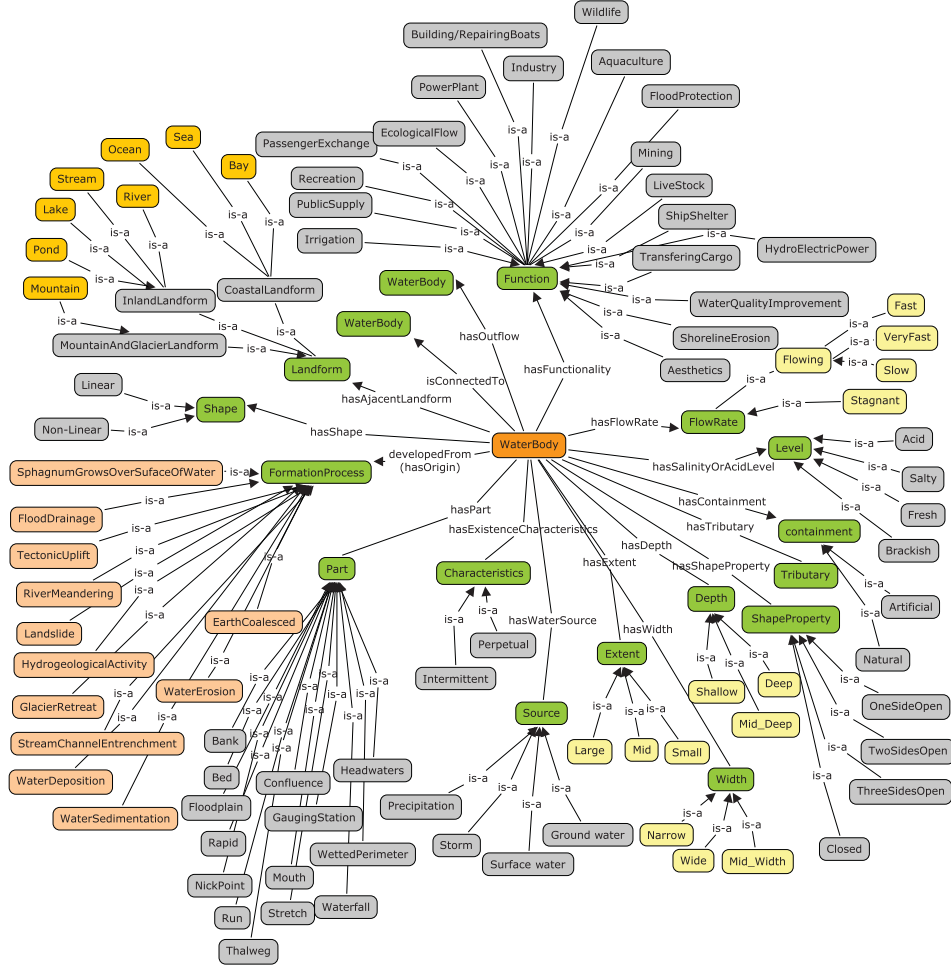


Figure 4. A feature space for “Waterbody”.

$A_k = \{x | x \in \text{Range}(t_{a_k}, f_i)\}, B_k = \{x | x \in \text{Range}(t_{b_k}, f_i)\}$: Range domain of feature i given objects t_{a_k} and t_{b_k} ;

$\text{Sim}(f_i, t_{a_k}, t_{b_k})$: The contribution of feature f_i to the similarity measure between t_{a_1} and t_{b_1} .

Basically, we can compute any feature’s contribution to the similarity measure of an object pair (t_{a_1}, t_{b_1}) using

$$\text{Sim}(f_i, t_{a_1}, t_{b_1}) = \frac{|A_1 \cap B_1|}{|A_1 \cup B_1|}. \quad (8)$$

As Equation (8) shows, the contribution of feature f_i in measuring the similarity between objects t_{a_1} and t_{b_1} is the ratio between shared members of A_1 and B_1 and the range they cover in total. So, the more members shared by A_1 and B_1 , the greater the contribution the feature f_i makes to the similarity between objects A_1 and B_1 . However, more rules need to be defined to handle special cases. The rules are as follows:

Table 1. Features used to describe a water body concept.

ID	Feature	Description
1	<i>hasFormationProcess</i>	The process through which a water body comes into existence, e.g. glacier retreat.
2	<i>hasPart</i>	The parts that compose a water body, e.g. mouth of a river.
3	<i>hasExistenceCharacteristic</i>	The continuous or periodical existence of water in a water body over time, e.g. ocean is perpetual.
4	<i>hasWaterSource</i>	Where is water in the water body from? E.g. precipitation.
5	<i>hasExtent</i>	The size of a non-linear water body, such as ocean.
6	<i>hasWidth</i>	The width of a linear water body, such as a river.
7	<i>hasDepth</i>	The depth of a water body.
8	<i>hasShape</i>	What shape is the water body in? E.g. a linear river and a non-linear lake.
9	<i>hasShapeProperties</i>	If the shape of water body is non-linear, what is the shape like? E.g. Gulf is a one-side open water body with three sides surrounded by land.
10	<i>hasTributary</i>	Whether a water body has one or more tributary. E.g. A mature river that flows slowly tends to have many tributaries.
11	<i>hasContainment</i>	Whether a water body is formed naturally or is human made.
12	<i>hasSalinityLevel</i>	The salinity of water, e.g. the water in ocean is salty.
13	<i>hasFlowRate</i>	The rate indicating how fast water flows in a water body, e.g. water in a lake is stagnant.
14	<i>hasFunction</i>	The functionality that a water body can be used for, e.g. irrigation.
15	<i>hasOutflow/hasConfluence</i>	Where the water flows to? E.g. rivers always flow into oceans.
16	<i>isConnectedTo</i>	The water body that a water body is connected with, e.g. a delta is always connected with sea or ocean.
17	<i>hasAdjacentLandform</i>	The landform of a water body is next to, e.g. an Arroyo is always found in mountainous region.

Rule I: For any object pairs (t_{a_1}, t_{b_1}) and (t_{a_2}, t_{b_2}) and a given feature f_i , if

$$(1) |A_1 - B_1| \cdot |B_1 - A_1| = 0 \text{ and}$$

$$(2) |A_2 - B_2| \cdot |B_2 - A_2| \neq 0$$

$$\text{Sim}(f_i, t_{a_1}, t_{b_1}) \geq \text{Sim}(f_i, t_{a_2}, t_{b_2}),$$

$$\text{when } \frac{|A_1 \cap B_1|}{|A_1 \cup B_1|} = \frac{|A_2 \cap B_2|}{|A_2 \cup B_2|} \neq \emptyset.$$

Rule II: For any object pairs (t_{a_1}, t_{b_1}) , (t_{a_2}, t_{b_2}) , (t_{a_3}, t_{b_3}) , if

$$(1) A_1 \cup B_1 \neq \emptyset \text{ and } |A_1| \cdot |B_1| \neq 0$$

$$(2) A_2 \cup B_2 = \emptyset$$

$$(3) A_3 \cup B_3 \neq \emptyset \text{ and } |A_3| \cdot |B_3| = 0$$

$$\text{Sim}(f_i, t_{a_1}, t_{b_1}) > \text{Sim}(f_i, t_{a_2}, t_{b_2}) > \text{Sim}(f_i, t_{a_3}, t_{b_3}).$$

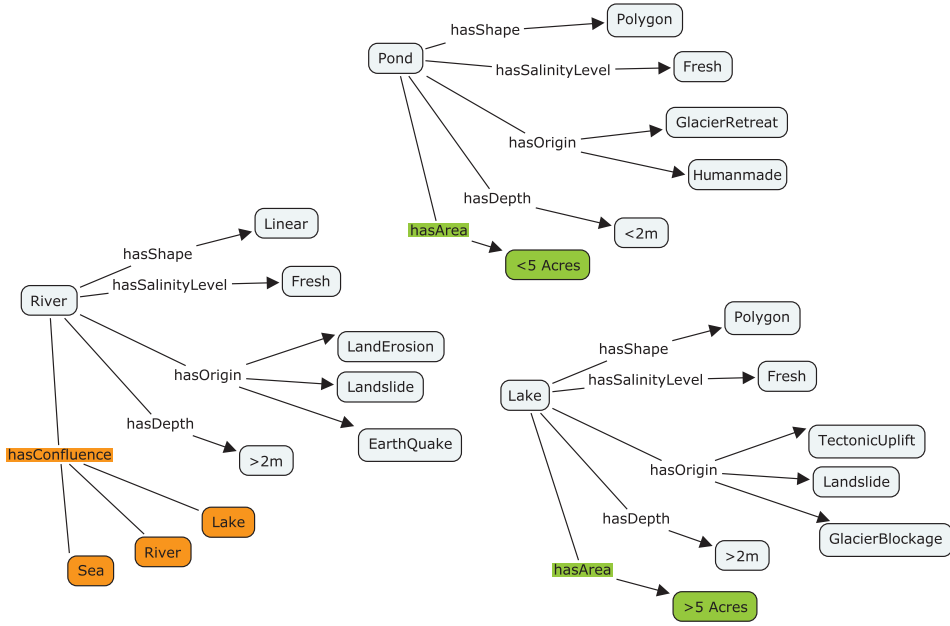


Figure 5. Semantic definition of three water body objects.

Case I in Table 2 shows the pairs that fit conditions (1) and (2) in Rule I respectively. For both case I (a) and case I (b), they have the same contribution to similarity given the same common and total features according to the definition in Equation (8). However, in case I (a), set B_1 is completely contained in set A_1 ; therefore, the contribution of a feature to objects in case I (a) should be larger than that in case II (b), as Rule I defined. Given this rule, the contribution to similarity can be computed as follows:

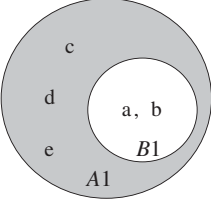
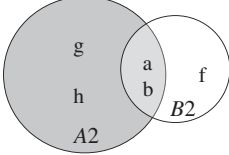
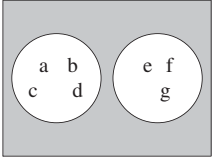
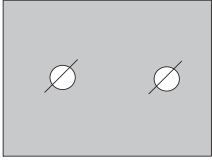
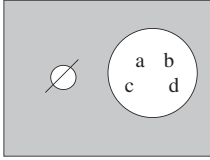
$$\text{Sim}(f_i, t_{a_1}, t_{b_1}) = \frac{|A_1 \cap B_1|}{|A_1 \cup B_1|}; \text{Sim}(f_i, t_{a_2}, t_{b_2}) = \beta \frac{|A_2 \cap B_2|}{|A_2 \cup B_2|} \quad (9)$$

where $\beta = \text{threshold} \in [0.9, 1]$

As an example, among the set of water body objects {River, Lake, Pond} in Figure 5, all have the same feature ‘hasShape’. According to other feature-based models, the contributions of the feature ‘hasShape’ are the same for each concept pair. But practically, the contribution of the above feature is larger in {Lake, Pond} pair than {Lake, River} and {Pond, River} pairs because both lake and pond have an oval shape, whereas a river is always linear. Thus, by considering the values in the range set of each common feature, this method can obtain more accurate values in similarity measurement than other feature matching models.

Case II in Table 2 shows the situation when the two range sets of a feature of two given objects have no common element. Case II (a) indicates that both objects share the feature, but does not have any intersection in the range set – condition 1 in Rule II; (b) indicates that neither of the objects has feature f_i – condition 2 in Rule II; (c) indicates that one object has feature f_i and the other does not – condition 3 in Rule II. Intuitively, the contribution of feature f_i to the similarity should be larger when this feature can be used to describe the objects than when neither of the objects has this feature, and slightly larger than when one

Table 2. Case examples for object pairs for Rules I and II.

 <p>(a)</p>  <p>(b)</p>	Case I
 <p>(a)</p>  <p>(b)</p>  <p>(c)</p>	Case II

object has the feature and the other does not, as Rule II defines. Given this rule, we define the contributions as follows:

$$\begin{aligned}
 \text{Sim}(f_i, t_{a_1}, t_{b_1}) &= \alpha \frac{1}{|A_1 \cup B_1|} \\
 \text{Sim}(f_i, t_{a_2}, t_{b_2}) &= \alpha \text{Sim}(f_i, t_{a_1}, t_{b_1}) \\
 \text{Sim}(f_i, t_{a_3}, t_{b_3}) &= \frac{|A_3 \cap B_3|}{|A_3 \cup B_3|} = 0
 \end{aligned} \tag{10}$$

where α is a tuning factor ($\alpha = 0.5$ for this case study). As an example, both water bodies ‘Pond’ and ‘River’ have the common feature ‘has Origin’. The range of this feature for ‘Pond’ is {GlacierRetreat, Manmade} and that for ‘River’ is {LandErosion, Landslide, Earthquake}. According to Rule II and Equation (10), the contribution of feature ‘has Origin’ to the similarity of {Pond, River} is 0.1 ($0.5(1/5)$) rather than 0, obtained directly from Equation (8).

3.4. Training process

Once rules for calculating contributions of both common and differential features are defined, the input pattern of the neural net can be mapped from pairs of objects and the similarity of the objects computed in Section 3.3. The neural net input includes a vector of multi-dimensional parameters and a known output result. Features are mapped onto the multi-dimensional parameters, and the value of each parameter is the contribution of the specific feature to the similarity of the two objects in pairs. The known output result is obtained from human ranking results on the sample data. Through an iterative training process of the designed neural network, the goal (similarity between the pairs of spatial

objects ranked by machine is highly correlated to human ranking) can be achieved with generated transition matrices.

Another issue for the similarity measurement is the timely update of a similarity matrix as the amount of knowledge (in this case, it is the water body ontology) increases. Once a new instance is populated into the ontology, its similarity with other instances in the ontology will be calculated automatically. Using the obtained transition matrix, a forward propagation can be conducted N (number of instances in ontology) times to calculate the missing similarity values. This achievement is based on the premise that the schema (object-level) of the ontological framework (recall Figure 4) is consistent, or the whole training process must be repeated for new transition matrices.

4. Assessing the ANN-based similarity measure approach

A frequently used experiment for assessing the semantic similarity is, to distribute to 38 undergraduate subjects 30 pairs of nouns that cover high, intermediate and low levels of similarity (Rubenstein and Goodenough 1965). In this study, the design of the experiments was slightly different from the one used by Rubenstein and Goodenough (1965) because the concepts measured are specialized for the hydrology domain, so measurements obtained from subjects who have little background in this domain may be biased because of the lack of domain knowledge.

Therefore, a new experiment was designed to satisfy the criterion mentioned above. The human subjects were asked to provide similarity scores for three groups of concept pairs. Subjects ranked the concept pairs from least to most similar, the lowest score was 0 for the least similar pair in a group whereas a score of 100 was assigned to the most similar pair. When scoring the similarity of one pair, the subject had to consider the relative distance of the similarity of this pair to that of the other pairs within the same group. The three groups are linear water body non-linear open water body I and non-linear water body II. According to the background of the human subjects (graduate students or hydrology experts), different surveys were given. The survey for the graduate subjects included 10 pairs of terms in each group. The survey designed for hydrology experts included all questions in the survey designed for graduate subjects, plus 33 other pairs. The extra pairs contained concepts from across groups, e.g. one is from the linear, whereas the other is from the non-linear water body group, e.g. (River, Lake), as shown in Table 3.

Based on the collected experimental data, the following assessment was conducted to evaluate the performance of ANN when enabling the automated similarity measurement as described in the following sections.

4.1. Quickness of convergence vs learning rate

The learning rate controls the speed of ANN learning by affecting the changes being made to the weights of transition matrices at each step. The performance of the ANN algorithm is very sensitive to the proper setting of the learning rate (Amini 2008). If the changes applied to the weights are too small, the algorithm will take too long to converge. If the changes are too large, the algorithm becomes unstable and oscillates around the error surface. This experiment determined the optimum network for automated similarity measurement through the result from this learning rate investigation. Here, 'optimum' is measured by the Mean Square Error (MSE) between the network outputs and the target outputs obtained from the human subject experiments. The initial parameters used for training the network are shown in Table 4. Parameter 1 is the largest number of steps that the ANN

Table 3. Survey conducted with human subjects.

Subject type	Survey A (linear)	Survey B (non-linear I)	Survey C (non-linear II)	Survey D (cross-group)
Both	%(River,Fjord)	%(Sea,Ocean)	%(Port,Dock)	%(Swamp,Marsh)
	%(River,Creek)	%(Sea,Bay)		%(Wetland,Swamp)
	%(River,Brook)	%(Sea,Gulf)		%(Wetland,Marsh)
	%(River,Bayou)	%(Sea,Cove)		%(Wetland,Bog)
	%(Creek,Fjord)	%(Sea,Harbor)		%(Wetland,Fen)
	%(Creek,Arroyo)	%(Sea,Port)		%(Swamp,Bog)
	%(Creek, Brook)	%(Sea,Dock)		%(Swamp,Fen)
	%(Brook,Arroyo)	%(Bay,Gulf)		% Bog,Fen)
	%(Creek, Bayou)	%(Bay,Cove)		
	%(Bayou,Brook)	%(Harbor,Port)		
		%(Harbor,Dock)		
		%(Port,Dock)		
Expert subject only	%(Bayou,Arroyo)	%(Ocean,Bay)	%(Wetland,Fen)	%(Lake,Arroyo)
	%(Bayou,Fjord)	%(Ocean,Gulf)	%(Marsh,Bog)	%(Lake,Bayou)
	%(Brook,Fjord)	%(Ocean,Cove)	%(Marsh,Fen)	%(Lake,Brook)
	%(Fjord,Arroyo)	%(Ocean,Harbor)		%(Lake,Creek)
	%(River,Arroyo)	%(Ocean,Port)		%(Lake,Fjord)
		%(Ocean,Dock)		%(Lake,River)
		%(Bay,Harbor)		%(Pond,Arroyo)
		%(Bay,Port)		%(Pond,Bayou)
		%(Bay,Dock)		%(Pond,Brook)
		%(Gulf,Cove)		%(Pond,Creek)
		%(Gulf,Harbor)		%(Pond,Fjord)
		%(Gulf,Port)		%(Pond,Lake)
		%(Gulf,Dock)		%(Pond,River)
		%(Cove,Harbor)		
		%(Cove,Port)		
		%(Cove,Dock)		

Table 4. Training parameters.

No.	Parameter	Value
1	Number of epochs	2000 5000 8000
2	Goal of performance function	10^{-3}
3	Initial learning rate	0.1
4	Training time	Infinity
5	Momentum coefficient	0.9

will run; Parameter 2 is measured by MSE; a value of 10^{-3} means that the ANN will stop training if $MSE < 10^{-3}$; Parameter 3 is the initial learning rate – in this experiment, the learning rate was set to different numbers in each training process; Parameter 4 sets the training expiration time to infinity. The introduction of Parameter 5 cut down the learning time and efficiently prevented the network from sticking at local optima.

Figure 6 shows the experimental results of neural network learning rate by the number of epochs. The *X*-axis indicates a learning rate ranging from 0.1 to 0.9 with interval 0.1, whereas the *Y*-axis indicates the number of ANN that must be trained until the result converges. As the network training uses heuristic techniques, it tends to become trapped in a local optimum because of the nature of the gradient descent algorithm from which these

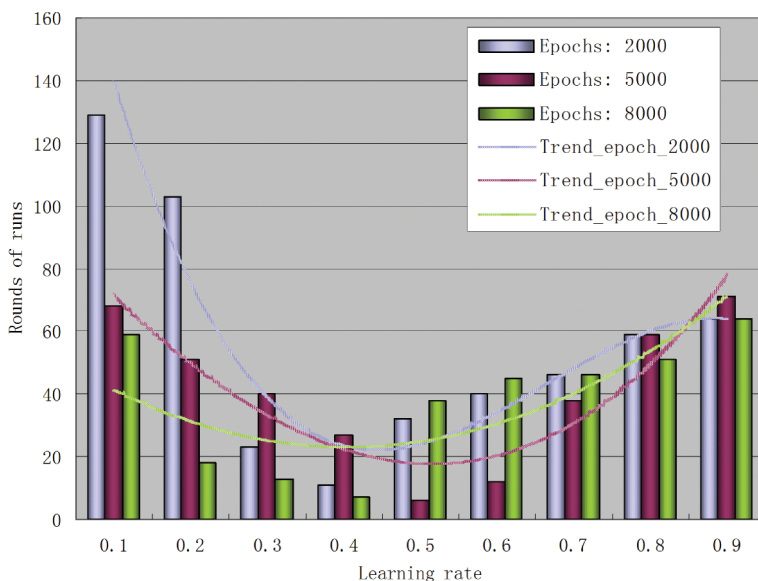


Figure 6. Number of training runs for the ANN needed in terms of various learning rates.

heuristic techniques were developed (Tan 2002). The strategy used here to compensate for the sticking problem was to retrain the network until the result achieved the performance function goal ($MSE < 10^{-3}$). The Y-axis records this number.

When the epoch is set to be large, the MSE between the training and target outputs is more likely to be within the tolerable range; therefore, the network needs fewer training runs. But the difference in complexity levels of different problems determines that the above assertion is not necessarily true. For the automated similarity measurement problem, the assertion is true only when the learning rate is less than 0.3. When the learning rate is more than 0.5, the setting of the epoch will not influence the number of training runs required. Another observation is that when the learning rate is less than 0.4, the number of training runs when the epoch equals 2000 decreases much faster than the decreasing rate of training runs when the epoch equals 5000 and 8000. This means that the designed neural network is most sensitive to change in learning rate when the epoch is set to 2000. Meanwhile, the trend curves in Figure 6 shows that for the same epoch of each training process, the necessary training runs decreases when the learning rate increases until the learning rate reaches 0.4. Based on the above analysis, when the epoch equals 2000 and the learning rate equals 0.4, the ANN performs the best and therefore those parameters were chosen for the following experiments.

4.2. Prediction accuracy vs number of hidden nodes

One great advantage of the ANN model is its ability to predict. Once experimental data are collected from human subjects, the neural network can be well trained. Using the trained network, the ANN model can provide automatic ranking for the pairs of concepts that are not ranked by humans. In order to accomplish this performance capacity, the experimental results from the human subjects were divided into two sets: 90% of the

results are considered as the testing set and the remaining 10% were considered as the validation set.

The correlation between the computational similarity models and human judgment has been widely used in previous studies to measure the accuracy of the model (Rada *et al.* 1989, Resnik 1999). The literature reports a correlation of 0.6 using a semantic-distance approach, 0.79 using an information-content approach and 0.83 using an extended-distance approach (Resnik 1999). In this article, a Pearson product-moment correlation coefficient r is used as one measure to investigate the association between the results from the trained ANN model and validation sets from human subjects.

$$r(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where X is the set of predicted similarities obtained from well-trained network and Y is the similarity values ranked by subjects.

The larger the r , the more accurate is the ANN model in predicting the similarity. The coefficient r only provides relevant qualified measurement of correlation for the two sets of data: the ANN generated set and the validation set. A higher correlation coefficient between the above datasets does not mean that the values in each corresponding pair are closer. A more accurate factor to measure the ‘prediction error’ is the square Root of MSE (RMSE) between values of each pair ranked by subjects and predicted by the ANN model:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\text{ANN}p_i - Hp_i)^2}{n}}$$

Therefore, the goal of an ANN model is to both maximize the coefficient r and minimize the RMSE. Figure 7 shows both Spearman coefficients (value in %) and the RMSE values. Of the nine hidden node settings of the ANN model, five result in high correlation (>85%) when making predictions. This means that the proposed ANN model is reliable in making predictions, and the high correlation shows that the ANN approach is better than most of the models previously considered. The best performance ($r = 94.86\%$, $\text{RMSE} = 11.47$) for the trained neural network occurs when hidden neuron equals 9.

As the number of hidden neurons determines the complexity of the neural network, although the ANNs with different neuron settings all satisfy the goal ($\text{RMSE} < 1$) when training the network, ANN will still cause overfitting or underfitting problems with too many or too few hidden neurons. According to Figure 7, when the number of hidden layers is 3 or 4, the network is not sufficiently complex and fails to detect fully the signal in the sample dataset. Therefore, this model leads to underfitting with low correlation and high RMSE in the prediction. When the number hidden neurons is equal to 10 or 11, the performance of the ANN declines, probably because the network experiences an overfitting problem where the noise is well fitted and this makes predictions not close enough to the range of the training data.

4.3. Accuracy of ANN prediction vs background of the subjects

This experiment examines the accuracy of ANN prediction, given the different backgrounds of the human subjects whose responses were the basis for the training and

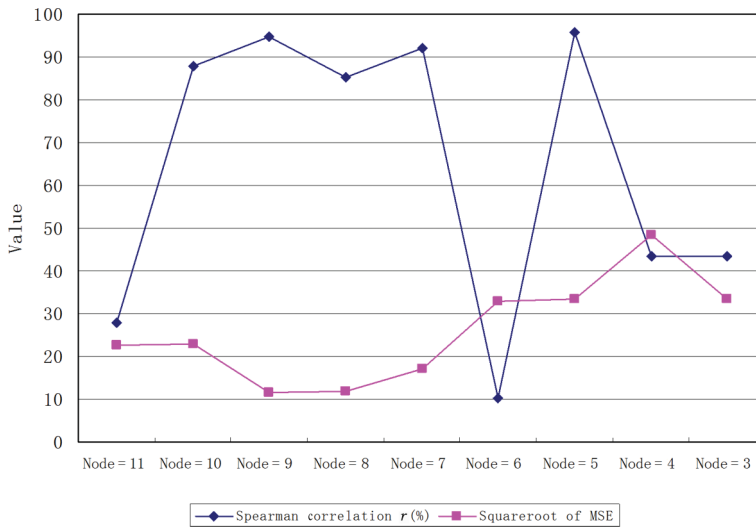


Figure 7. Prediction accuracy as a function of number of hidden nodes.

validation datasets. The ANN was trained with the optimal learning rate (0.4) and the optimal hidden neurons (9) obtained from the above experiments. As the sample dataset from human subjects for each group of pairs of water body objects was relatively small (refer to the number of pairs in Table 3), ranking data from other groups were borrowed to make sure that the total number of samples for the ANN training group was equal to or more than the number of total features ($\text{Num}(\text{feature}) = 17$) of the water body objects. Three pairs of the water body objects in each group were used for validation, and the rest of the pairs were used for training. As Table 5 shows, the correlation coefficients acquired for both types of subjects are all above 70% and still lower than the number acquired in Section 4.2, which was conducted on a larger sample. The ANN trained by data collected from graduate subjects has a lower correlation coefficient than that collected from the expert subjects, meaning that more noisy data exist in the survey of graduate subjects. Tracking their similarity rankings suggests two primary reasons for the increased noise: (a) graduate subjects tend to rank the similarity based more on the familiarity of the spatial concepts rather than the actual meaning; for example, ‘Bog’ and ‘Fen’ have many common features including shape, size, how they are developed, water source and water salinity. But most graduate subjects gave these pairs a low rank because of unfamiliarity with these water body terms. (b) Misunderstanding of the spatial concepts leads to much bias of the ranking results. In comparison, ranking results from hydrology experts is more reliable. From this experiment, we can conclude that collecting enough (>3 times of total feature sets in our case) reliable sample data is important for the ANN model to perform accurate prediction.

5. Application

The ESIP (Earth Science Information Partnership) semantic search testbed (Li 2010) aims to completely utilize the knowledge encoded in the ontologies and to incorporate the semantic similarity introduced in this article to provide a better search experience to users in the GIScience community. Several modules were developed to make the knowledge discovery more intelligent: semantic registration, semantic search and

Table 5. Accuracy of the ANN prediction for graduate subjects and expert subjects


Subject type \ Correlation	Non-linear		
	Linear	Non-linear I	Non-linear II
Graduate subject	0.71	0.82	0.79
Expert subject	0.81	0.85	0.91

geo-bridge. Semantic registration allows the collaborative population of domain knowledge through an online interface. This module facilitates the ontology development process, such as developing the feature-based water body ontology needed in this study. The semantic search module extends the search by providing all relevant terms in the query (Li *et al.* 2009). Geo-bridge links the semantic search client to popular geospatial web catalogues, including GOS (Geospatial One Stop), GCMD (Global Climate Master Directory), NCDC (National Climatic Data Center) and ECHO (Earth Observation ClearingHouse) (Li *et al.* 2008). Services discovered by the crawler (Li *et al.* 2010) are registered in the GOS portal and are made available for the semantic search client. The similarity matrix obtained from the neural network training described in this article was integrated to rank the relevance of recommended terms to the input keyword (lower right column in Figure 8). This mechanism enables the relevancy recommendation to users, especially those with limited domain knowledge, to refine the search and find the most appropriate dataset.

6. Conclusion and discussion

This article introduced a novel feature-based approach that utilized ANN to best simulate the human similarity ranking process based on training artificial neurons with sample data collected from human subjects. The collection and ontological modelling of spatial objects, the calculation of contribution for each feature of any two spatial objects and the ANN design were introduced. In several experiments, the ANN-based approach achieved good performance in terms of both correlation and mean standard error when measuring the similarity between ANN prediction and human subject rankings. Finally, an ESIP semantic search application that incorporates similarity based ranking was described. The similarity measurement provides an effective way of term disambiguation and intelligent query answering. The ESIP semantic web testbed incorporating this feature is capable of answering queries such as ‘What is the most similar water body concept to a River?’ and ‘What term can be used as an alternative to River when a user conducts a search?’ with the assistance of the similarity matrix obtained from this study.

Compared to existing methods, the proposed method enables similarity measurement on top of the advanced ontological framework based on DL. It is capable of handling knowledge represented with more constraints and interrelations and those organized in an interconnected graph besides a simple hierarchical tree. Meanwhile, the feature-based modelling of water body ontology provides a comprehensive knowledge set to describe and distinguish spatial features that are required to emulate human behaviour by machines. In addition, the ANN-based machine learning algorithm provides a global optimized function to complement other approaches in automating and precisely simulating the human perception process of similarity. We have also been successfully applying the theoretical studies to practical applications for more effective retrieval of geospatial data. Another unique



river

Search

Web

Number of results: 18 for river [definition].

Title: Canadian Rivers Data Set [ARCSS019]

Source: gcmd

OnlineAccessible: yes

Description: 2. Floods of May 30 to June 15, 2008, in the Iowa River and Cedar River Basins, Eastern Iowa [USGS OFR 2010.1190] River and Cedar River Basins. The storms were part of an exceptionally wet period from May 29... at rain gages in Iowa Falls and Clutier were 14.00 and 13.83 inches, respectively. Within the Iowa River

data

Title: North Shore Volunteer River Herring Counts [gcmd_170]

Source: gcmd

OnlineAccessible: yes

Description: herring in the Great Marsh and the North Coastal Watershed, Essex CountyMassachusetts. Monitoring began with the Parker River in 1997 and now includes

data

Title: Restoration and Establishment of Sea Run Fisheries [gcmd_...

Source: gcmd

OnlineAccessible: yes

Description: American shad and river herring to selected river... is renewedevery five years. Obtain additional adult river herring brood stock totransplant into restored

data

Title: Maryland Department of Natural Resources (MDNR) Fall Oyst...

Resources

☒ NOAA NCDC - 0

☒ NASA GCMD - 15

☒ FGDC GOS - 0

☒ NASA ECHO - 3

☒ NASA ESG - 0

Refine Your Search Here:

+Type1

+river

☐ Meander

Related Terms

+Earthrealm

☐ Fjord 80%

☐ Creek 75%

☐ Brook 70%

☐ Arroyo 70%

☐ Bayou 65%

☐ Lake 50%

☐ Pond 30%

Figure 8. Screenshot of ESI semantic search testbed.

feature of the proposed method is that the formulation of rules to calculate the contribution of a specific feature to the similarity of two spatial objects makes the context-aware similarity measure possible. As an example, in Section 3.3, we computed the contribution of the feature ‘hasShape’ to the similarity measure of {Lake, Pond}, {Lake, River} and {Pond, River}. By applying the rules, we obtained that the contribution of the above feature is larger in the {Lake, Pond} pair than the {Lake, River} and {Pond, River} pairs because both lake and pond have an oval shape, whereas a river is always linear. This capability could well answer context-aware questions such as ‘Which two water body features are more similar in terms of shape?’ Similarly, we will also be able to answer questions in other contexts, such as ‘find the most similar terms to ocean in terms of the salinity level’.

Several potential research directions attract our attention. As a pilot study to validate the feasibility of this proposed approach, the scale of knowledge encoded in the current water body ontology is relatively small. The concepts defined are at the object-level (e.g. ‘River’, ‘Lake’) rather than the instance level (e.g. Mississippi river or Lake Manasarovar). The similarity measure on the instantiations of water body concepts can answer more queries, such as ‘Which river is most similar to the Mississippi River?’ In addition to the queries a search engine can answer currently, the features defined to describe water body concepts will also allow a search engine to answer extended queries such as ‘Which river is most similar to the Mississippi River in terms of origin or functionality or geographical location?’ To implement this capability, we must extend the current ontological framework to include instance-level water body concepts as the next step in this research. In addition, knowledge from other science domains, such as geology, biology and astronomy, will also be modelled to validate and promote the ubiquity of the proposed methodology. Furthermore, training of the neural net is a time-consuming process, especially when the training set is large and the underlying pattern is complex. Therefore, parallelizing the algorithm to improve its efficiency for pre-processing is another issue to be studied.

Acknowledgement

This work was funded by the Earth Science Information Partnership (ESIP) Product and Service Committee and the Federal Geographic Data Committee (FGDC). The authors are grateful to Dr. Chaowei Yang, George Mason University, for providing his valuable comments on this research. The authors also thank Mr. Steve McClure for proofreading the article.

References

- Amini, J., 2008. Optimum learning rate in back-propagation neural network for classification of satellite images (irs-1d). *Scientia Iranica*, 15, 558–567.
- Berners-Lee, T., Hendler, J., and Lassila, O., 2001. The semantic web – a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284, 28–37.
- Budanitsky, A. and Hirst, G., 2001. Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. *Proceedings of the NACCL 2001 Workshop: on WordNet and other lexical resources: applications, extensions, and customizations*, 2–7 June, Pittsburgh, USA, 29–34.
- Cilibrasi, R. and Vitanyi, P., 2006. Similarity of objects and the meaning of words. *Proceedings of conference on theory and applications of models of computation*, 15–20 May, Beijing, China, 3959, 21–45.
- d’Amato, C., Staab, S., and Fanin, C., 2008. On the influence of description logics ontologies on conceptual similarity. In: A. Gangemi and J. Euzenat, eds., *Proceedings of the 16th knowledge engineering conference, EKAW2008. Vol. 5268, LNAI*. Berlin, Heidelberg: Springer-Verlag, 5268, 48–63.

- Eshera, M.A. and Fu, K.S., 1984. A graph distance measure for image-analysis. *IEEE Transactions on Systems Man and Cybernetics*, 14, 398–408.
- Gentner, D. and Markman, A.B., 1995. Similarity is like analogy: structural alignment in comparison. In: C. Cacciari, ed., *Similarity in language, thought and perception*, Brussels: BREPOLs, 111–147.
- Goldstone, R.L. and Son, J., 2005. Similarity. In: K.J. Holyoak & R.G. Morrison, eds., *Cambridge handbook of thinking and reasoning*, Cambridge: Cambridge University Press, 13–36.
- Janowicz, A., Raubal, M., and Kuhn, W., 2011. The semantic similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2, 29–57.
- Janowicz, K., 2006. Sim-dl: towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval. In: R. Meersman, Z. Tari, and P. Herrero, eds. *On the move to meaningful internet systems. Proceedings. Part II. Vol. 4278, LNCS 4278*. Berlin Heidelberg: Springer-Verlag, 1681–1692.
- Kebler, C., 2007. Similarity measurement in context. In: *Proceedings of the 6th international and interdisciplinary conference on Modeling and using context*, 20–24 August 2007, Roskilde, Denmark, 4635, 277–290.
- Li, W., 2010. *Automated data discovery, reasoning and ranking in support of building an intelligent geospatial search engine*. Dissertation (PhD). George Mason University, 168pp.
- Li, W., Yang, C., and Raskin, R., 2008. A semantic enhanced model for searching in spatial web portals. *Proceedings of semantic scientific knowledge integration AAAI/SSKI symposium*, Stanford University, Palo Alto, CA, 47–50.
- Li, W., Yang, C., and Raskin, R., 2009. A semantic-enabled meta-catalogue for intelligent geospatial information discovery. In: *Geoinformatics, 2009 17th International Conference*, 12–14 August, Fairfax, Virginia, 1–5.
- Li, W., Yang, C., and Yang, C., 2010. An active crawler for discovering geospatial web services and their distribution pattern – a case study of OGC web map service. *International Journal of Geographical Information Science*, 24, 1127–1147.
- Lin, D., 1998. An information-theoretic definition of similarity. In: J.W. Shavlik, ed., *Proceedings of the fifteenth international conference on machine learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc, 296–304.
- Losif, E. and Potamianos, A., 2007. Unsupervised semantic similarity computation using web search engines. In: *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*. Washington, DC: IEEE Computer Society, 381–387.
- Medin, D.L., Goldstone, R.L., and Gentner, D., 1993. Respects for similarity. *Psychological Review*, 100, 254–278.
- Rada, R., et al., 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19, 17–30.
- Raskin, R. and Pan M., 2005. Knowledge representation in the semantic Web for Earth and environmental terminology (SWEET). *Computer & Geosciences*, 31(9), 1119–1125.
- Resnik, P., 1999. 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.
- Richardson, R., Smeaton, A.F., and Murphy, J., 1994. Using wordnet as a knowledge base for measuring semantic similarity between words. In: *Proceedings of AICS conference*. Berlin, Heidelberg: Springer-Verlag.
- Rodriguez, M.A. and Egenhofer, M.J., 2003. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15, 442–456.
- Rodriguez, M.A. and Egenhofer, M.J., 2004. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, 18, 229–256.
- Rubenstein, H. and Goodenough, J.B., 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8, 627–637.
- Santos, P., Bennett, B., and Sakellariou, G., 2005. Supervalue semantics for an inland water feature ontology. In: F. Giunchiglia, ed. *Proceedings of the 19th international joint conference on Artificial intelligence*. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc, 564–569.
- Sattath, S. and Tversky, A., 1987. On the relation between common and distinctive feature models. *Psychological Review*, 94, 16–22.

- Schwering, A., 2008. Approaches to semantic similarity measurement for geospatial data: a survey. *Transactions in GIS*, 12, 5–29.
- Schwering, T. and Kuhn, W., 2009. A hybrid semantic similarity measure for spatial information retrieval. *Spatial Cognition and Computation*, 9, 30–63.
- Seco, N., Veale, T., and Hayes, J., 2004. An intrinsic information content metric for semantic similarity in wordnet. In: R.L. de Mántaras and L. Saitta, eds., *Proceedings of 16th European conference on artificial intelligence*. Amsterdam, Netherland: IOS Press, 1089–1090.
- Singh, Y. and Chauhan, A.S., 2005. Neural network in data mining. *Journal of Theoretical and Applied Information Technology*, 5, 37–42.
- Sneth, A.P., 1999. Changing focus on interoperability in information systems: from system, syntax, structure to semantics. In: M. Goodchild, *et al.*, eds. *Interoperating geographic information systems*. Norwell, MA: Kluwer Academic Publishers, 1–25.
- Tan, Y., 2002. A neural network approach for signal detection in digital communication. *The Journal of VLSI Signal Processing*, 32, 45–54.
- Tversky, A., 1977. Features of similarity. *Psychological Review*, 84, 327–352.
- Tversky, A. and Gati, I., 1978. Studies of similarity. In: E. Rosch and B. Lloyd, eds. *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum, 79–98.