Semantic Similarity Measurement Based on Knowledge Mining: An Artificial Neural Net Approach

Journal:	*International Journal of Geographical Information Science*
Manuscript ID:	Draft
Manuscript Type:	Research Article
Keywords:	Geographic Information Science < Keywords Relating to Theory, Semantics < Keywords Relating to Theory

Semantic Similarity Measurement Based on Knowledge Mining: An Artificial Neural Net Approach

Abstract

This paper 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 modeled by a set of features that best represent the spatial, temporal and descriptive attributes of the concepts, such as origin, shape, function and etc. A 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.

Keyword: Neural Network, Semantic Similarity, Search Engine, Ranking, Ontology, Logic

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 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, while 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, etc. Semantic similarity measures how closely two concepts are related by providing a ***quantitative*** value. The objective of paper is to apply a 'machine expert' to simulate the human perception process in measuring the semantic similarity between spatial objects quantitatively. The hydrology domain will be the context of this study.

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 (Cilibrisi and Vitanya, 2006; Janowicz, 2006; Janowicz et al. 2011). Due to 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 (Sheth, 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 05/10/11 had received in total 289,000,000 questions asking about the difference from one concept to another and Microsoft Bing search by 05/10/11 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 modeling and similarity identification between geometric characteristics of a single object and geographic relationships between spatial objects help 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; Iosif 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 will start by formalizing the problem after presenting a use case, and then discuss the build-up of ontological data and the methodology in use, and at last the experiments conducted to validate the proposed methodology.

2. A Use Case

FIGURE 1 NEAR HERE

Conceptual vagueness is a ubiquitous problem in the geospatial domain. Santos (Santos et al. 2005) demonstrated an example of vagueness in water features, as Figure 1 shows. This image 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 to these three research questions are the focus of this paper.

3. 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; Resnik, 1999; Lin, 1998; Seco et al. 2004), and feature matching models (Tversky, 1997; Tversky and Gati, 1978; Sattath and Tversky, 1987; Rodriguez and Egenhofer, 2003; Rodriguez and Egenhofer, 2004).

Edge-counting techniques are based on a network of semantic relations between concepts, and involve calculating 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 categroies 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 judgment.

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 are still based on a knowledge base with a simple logic is not suitable for mainstream knowledge representation, such as First Order Logic (FOL) and Description Logic (DL). This is because in the DL, there are more constraints of an object and interrelations between objects than that in taxonomy. And the structure of knowledge base changes from a tree-based structure to an interconnected graph. Therefore, 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 since it would be too time-consuming and inflexible. Instead, there is a need of 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.

4. Problem Definition

Similarity measurement can be formally described as follows: from a collection of interrelated objects D , find a subset S of terminologies such that the similarity between each element s_i of S and the given terminology t_g ranked by machine is highly correlated to human ranking. A mathematical description of this problem is as follows:

$$\text{For } \forall s_i,\ \Gamma(t_g, s_i) \rightarrow \mathrm{H}(t_g, s_i)$$

Where $i \le |S| \le |D|$, ($|X|$ equals the size of set X). H is a function for human ranking and Γ is the goal function to achieve.

To expand this measurement to the scale of a whole dataset D, $\Gamma(d_i, d_j)$ is calculated for $\forall d_i, d_j \in D,\ 0 \le i, j \le |D|$ and $i \ne j$, by which means the distribution space for range of Γ could be obtained, as shown in Figure 2.

FIGURE 2. NEAR HERE

5 Proposed Methodology

Artificial Neural Networks (ANN), 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 (for example, 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 since 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 paper, MLFFN is utilized to measure semantic similarity automatically.

5.1 MLFNN Algorithm

A MLFNN is used here to conduct numerical learning to simulate the knowledge propagation in a biological neural network. Figure 3 illustrates a general design of the multi-layer neural net which has multiple inputs and outputs:

FIGURE 3 NEAR HERE

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):

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

The detailed algorithm is:

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:

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

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

$$W = W + \Delta W \tag{4}$$

$$V = V + \Delta V \tag{5}$$

    b. perform forward propagation as follows:

$$H_j = \sigma(\sum_k V_{jk} I_k) \tag{6}$$

$$O = \sigma(\sum_j W_{ij} H_j) \tag{7}$$

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

c. Calculate the mean squared error between each output and desired output. If the worst error is lower than a given good-minimum-error, the network is finished 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.

5.2 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 of significant importance 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, description logic (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 behavior, 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 paper, 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 4 shows all the water body concepts that are modeled 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 modeled but not yet considered in measuring the similarity with other objects in the current phase because there is very limited information about them encoded in the sources of water body ontologies/articles.

FIGURE 4 NEAR HERE

Figure 5 shows the ontological framework of water body described by 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 5 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, TransferingCargo, WaterQualityImprovement, ShorelineErosion, and Aesthetics}. Each member in this set can be used to replace “Function” in the triple expression {“WaterBody,” “hasFuntionality,” Function}. For example, a triple {“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.

FIGURE 5 NEAR HERE

TABLE 1 NEAR HERE

Figure 6 demonstrates semantic definitions of three different water body objects: river, lake and pond. All of them share the feature set {hasShape, hasSalinityLevel, hasOrigin, hasDepth}, although the range of the features are different. Feature nodes in orange are the ones only owned by "River;" nodes in green are the ones shared by “Pond” and “Lake” but not “River”. By modeling each water body object, the similarity can be measured more precisely, from the granularity of a set of features (such as what is adopted in the MSDM model) down to a single feature. The following section defines the rules for determining the contribution of a single feature to the similarity measurement of two or more spatial objects.

FIGURE 6 NEAR HERE

5.3 Derivation of training data from the knowledge base and rules

As a neural net is a numerical learning algorithm, it requires numerical input parameters. Therefore, there is a need to numerically express the contribution of each feature toward measuring the similarity of two or more spatial objects. To formulate the rules for calculating the contribution, consider the following parameters:

k : Index of spatial objects of different kinds;

$(t_a,\ t_b)$: Pair of spatial objects for similarity measure;

f_i: One of the 17 features defined in Table 1, i is the index;

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

$Sim(f_i, t_{a_k}, t_{b_k})$: 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 $(t_{a_1},\ t_{b_1})$ by:

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

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 of the same member shared by A_1 and B_1, the greater the contribution the feature makes to the similarity between the two objects. However, more rules need to be defined to handle special cases. The rules are:

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) $|A_1 - B_1| \cdot |B_1 - A_1| = 0$, and

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

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

When $\dfrac{|A_1 \cap B_1|}{|A_1 \cup B_1|} = \dfrac{|A_2 \cap B_2|}{|A_2 \cup B_2|} \neq \varnothing$;

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) $A_1 \bigcup B_1 \neq \varnothing$ and $|A_1|\cdot|B_1| \neq 0$; and

(2) $A_2 \bigcup B_2 = \varnothing$; and

(3) $A_3 \bigcup B_3 \neq \varnothing$ and $|A_3|\cdot|B_3| = 0$

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

TABLE 2 NEAR HERE

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 B1 is completely contained in set A1, 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:

$$Sim(f_i, t_{a_1}, t_{b_1}) = \frac{|A_1 \bigcap B_1|}{|A_1 \bigcup B_1|};\ Sim(f_i, t_{a_2}, t_{b_2}) = \beta \frac{|A_2 \bigcap B_2|}{|A_2 \bigcup B_2|} \qquad (9)$$

Where $\beta = threthhold \in [0.9, 1)$

As an example, among the set of water body objects {River, Lake, Pond} in Figure 6, 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} because lake and pond both have an oval shape while 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 not having 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 object has the feature and the other does not, as Rule II defines. Given this rule, we define the contributions as:

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

where α is a tuning factor ($\alpha = 0.5$ for this case study). As an example, both water bodies "Pond" and "River" have common feature "hasOrigin." 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 "hasOrigin" to the similarity of {Pond, River} is 0.1 (0.5*(1/5)) rather than 0, obtained directly from equation (8).

5.4 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 5.3. 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 sample data. Through the process listed in Figure 7, the goal function $\Gamma(d_i, d_j)$ introduced in Section 4 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 missing similarity values. This achievement is based on the premise that the schema (object-level) of the ontological framework (recall Figure 5) is consistent, or the whole training process must be repeated for new transition matrices.

FIGURE 7 NEAR HERE

6 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 experiments was slightly different than 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 due to 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 while 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 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 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, while the other is from the non-linear water body group, e.g. (River, Lake), as shown in Table 3.

TABLE 3 NEAR HERE

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.

6.1 Quickness of Convergence v.s. 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 will run; Parameter 2 is measured by MSE; a value of 10^-3 means 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 cuts down the learning time and efficiently prevented the network from sticking at local optima.

TABLE 4 NEAR HERE

Figure 8 shows the neural network learning rate experimental results by the number of epochs. The X-axis indicates a learning rate ranging from 0.1 to 0.9 with interval 0.1, while 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 due to the nature of the gradient descent algorithm from which these 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.

FIGURE 8 NEAR HERE

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 for epoch = 2000 decreases much faster than the decreasing rate of training runs for epoch = 5000 and epoch =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 8 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.

6.2 Prediction Accuracy v.s. 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 (Resnik, 1999; Rada et al. 1989). 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 paper, 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.

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

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

The larger r is, the more accurate the ANN model is 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:

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

Therefore, the goal of an ANN model is to both maximize the coefficient r and minimize the RMSE. Figure 9 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%, 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 (RMSE<1) when training the network, ANN will still cause overfitting or underfitting problems with too much or too few hidden neurons. According to Figure 9, 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 hidden neurons are equal to 10 or 11, the performance of the ANN declined, 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.

FIGURE 9 NEAR HERE

6.3 Accuracy of ANN Prediction v.s. 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 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 is relatively small (refer to the number of pairs in Table 3), ranking data from other groups are 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 (Num(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 6.2, 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 exists 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 due to 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.

TABLE 5 NEAR HERE

7. Application

The ESIP semantic search testbed (Li and Yang, 2009; Li, 2010) aims to completely utilize the knowledge encoded in the ontologies and to incorporate the semantic similarity introduced in this paper 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 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. 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). Services discovered by the crawler (Li et al. 2010) are registered into the GOS portal and are made available for the semantic search client. The similarity matrix obtained from the neural network training described in this paper was integrated to rank the relevance of recommended terms to the input keyword (lower right column in Figure 10). This mechanism enables the relevancy recommendation to uses, especially those with limited domain knowledge, to refine the search and find the most appropriate dataset.

FIGURE 10 NEAR HERE

8. Conclusion and Discussion

This paper introduces a novel feature-based approach that utilizes 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 modeling of spatial objects, the calculation of contribution for each feature of any two spatial objects, and the ANN design are 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. In the conclusion, 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 able to answer 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.

In the future, several 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 Manasarover). 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 a next step in the research.

In addition, the knowledge from other science domains, such as geology, biology, and astronomy will also be modeled 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, how to parallelize the algorithm to improve its efficiency for pre-processing is another issue to be studied.

References

Amini, J. (2008) Optimum learning rate in back-propogation neural network for classification of satellite images (irs-1d). *Scientia Iranica* **15** , pp. 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** , pp. 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.*, pp. 29-34.

Cilibrasi, R. and Vitanyi, P. (2006) Similarity of objects and the meaning of words. *Theory and Applications of Models of Computation, Proceedings* **3959** , pp. 21-45.

d'Amato, C., Staab, S. and Fanin, C. (2008) On the influence of description logics ontologies on conceptual similarity. In *Lnai 5268*, pp. 48-63 (Berlin Heidelberg: Springer-Verlag).

Eshera, M.A. and Fu, K.S. (1984) A graph distance measure for image-analysis. *Ieee Transactions on Systems Man and Cybernetics* **14** , pp. 398-408.

Gentner, D. and Markman, A.B. (1995) Similarity is like analogy: Structural alignment in comparison. *Similarity in Language, Thought and Perception*, pp. 111-147.

Goldstone, R.L. and Son, J. (2005) Similarity. *Cambridge Handbook of Thinking and Reasoning*, pp. 13-36.

Janowicz, A., Raubal, M. and Kuhn, W. (2011) The semantic similarity in geographic informaiton retrieval. *Journal of Spatial Information Science*, 2011, pp. 29-57.

Janowicz, K. (2006) Sim-dl: Towards a semantic similarity measurement theory for the description logic alcnr in geographic information retrieval. In *LNCS 4278*, R. Meersman, Z. Tari and H. P. (Eds.) (Berlin Heidelberg: Springer-Verlag), pp. 1681-1692.

Kebler, C. (2007) Similarity measurement in context. *Proceedings of the 6th international and interdisciplinary conference on Modeling and using context* **4635** , pp. 277-290.

Li, W. (2010) Automated data discovery, reasoning and ranking in support of building an intelligent geospatial search engine. Ph.D. Dissertation, George Mason University, 168pp.

Li, W., Yang, C. and Raskin, R. (2009) A semantic-enabled meta-catalogue for intelligent geospatial information discovery. In *Geoinformatics, 2009 17th International Conference on*, pp. 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** , pp. 1127-1147.

Lin, D. (1998) An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 296-304 (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.).

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*, pp. 381-387 (Washington D.C., USA: IEEE Computer Society).

Medin, D.L., Goldstone, R.L. and Gentner, D. (1993) Respects for similarity. *Psychological Review* **100** , pp. 254-278.

Rada, R., Mili, H., Bicknell, E. and Blettner, M. (1989) Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on* **19** , pp. 17-30.

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** , pp. 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** , pp. 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** , pp. 229-256.

Rubenstein, H. and Goodenough, J.B. (1965) Contextual correlates of synonymy. *Communications of the ACM* **8** , pp. 627-&.

Santos, P., Bennett, B. and Sakellariou, G. (2005) Supervaluation semantics for an inland water feature ontology. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pp. 564-569 (Edinburgh, Scotland: Morgan Kaufmann Publishers Inc.).

Sattath, S. and Tversky, A. (1987) On the relation between common and distinctive feature models. *Psychological Review* **94** , pp. 16-22.

Schwering, A. (2008) Approaches to semantic similarity measurement for geospatial data: A survey. *Transactions in GIS* **12** , pp. 5-29.

Schwering, T. and Kuhn, W. (2009) A hybrid semantic similarity measure for spatial information retrieval. *Spatial Cognition and Computation* **9** , pp. 30-63.

Seco, N., Veale, T. and Hayes, J. (2004) An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of 16th European Conference on Artificial Intelligence* (Amsterdam, Netherland: IOS Press), pp. 1089-1090.

Singh, Y. and Chauhan, A.S. (2005) Neural network in data mining. *Journal of Theoretical and Applied Information Technology*, **5**, pp. 37-42.

Sneth, A.P. (1999) Changing focus on interoperability in information systems: From system, syntax, structure to semantics. In *Interoperating geographic information systems*, M. Goodchild, M.J. Egenhofer, R. Fegeas and C. Kottman (Eds.) (Norwell, MA, USA: Kluwer Academic Publishers).

Tversky, A. (1977) Features of similarity. *Psychological Review*, **84**, pp. 327-352.

Tversky, A. and Gati, I. (1978) Studies of similarity. In *Cognition and categorization*, E. Rosch and B. Lloyd (Eds.), pp. 79-98 (Hillsdale, NJ: Lawrence Erlbaum).

Tan, Y. (2002) A neural network approach for signal detection in digital communication. *The Journal of VLSI Signal Processing* **32** , pp. 45-54.

List of Figures

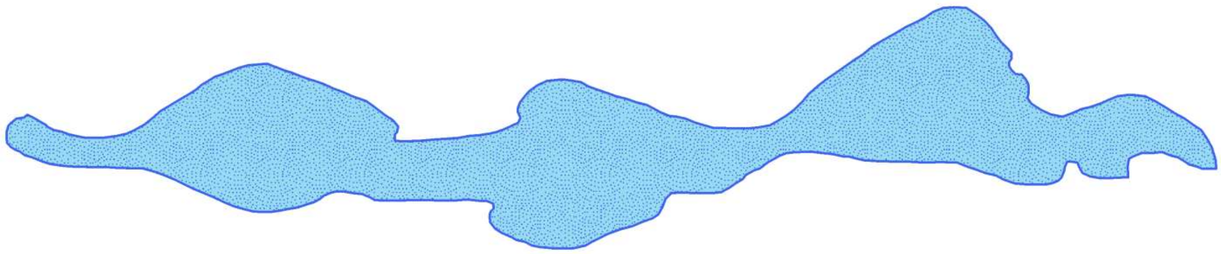


Figure 1 Vagueness in Water Features: Three Lakes or a Meandering River? (Santos et al. 2005)

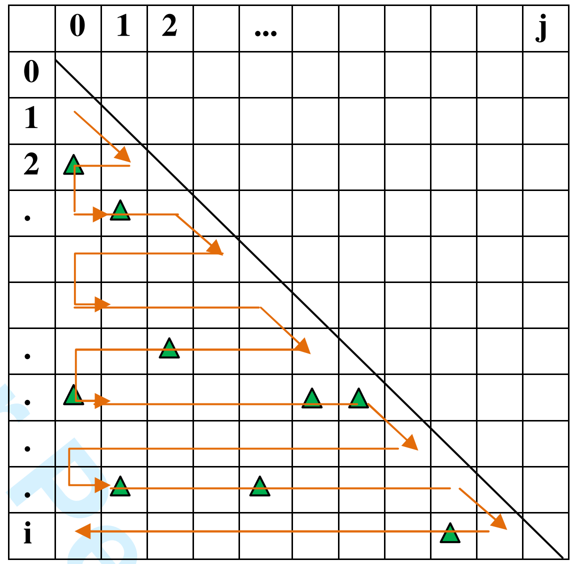


Figure 2 Distribution space for Γ. Lower triangle of the matrix filled with arrows demonstrates $\Gamma(d_i, d_j)$ and the green triangles in cells are the example training datasets available as training dataset in the methodology discussed in the following section. The yellow line means the coverage of all similarity measures from the training datasets.

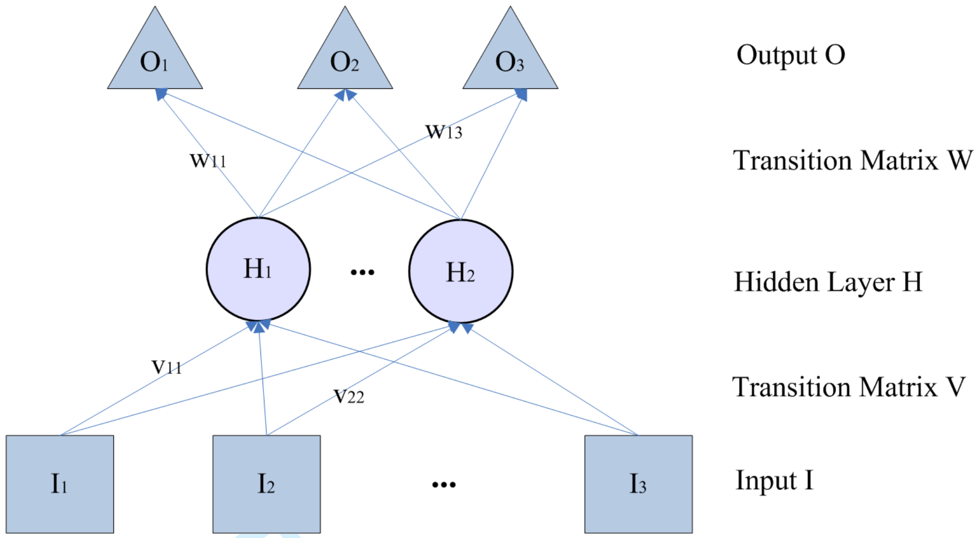


Figure 3 Design of a MLFNN.

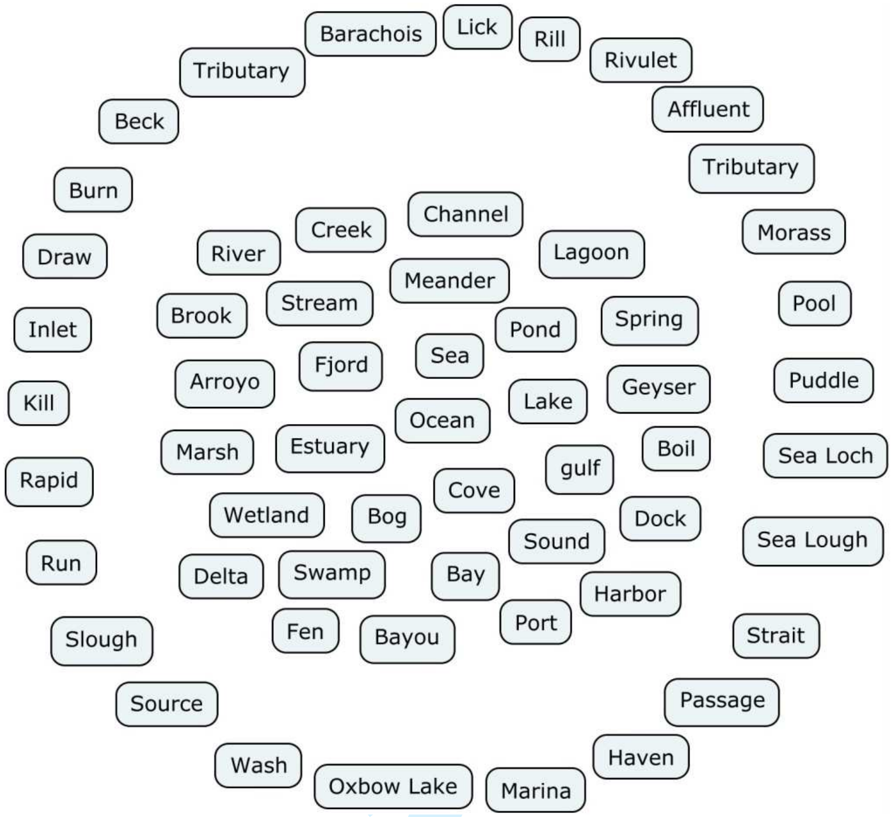


Figure 4 Water Body Concepts Used for Training.

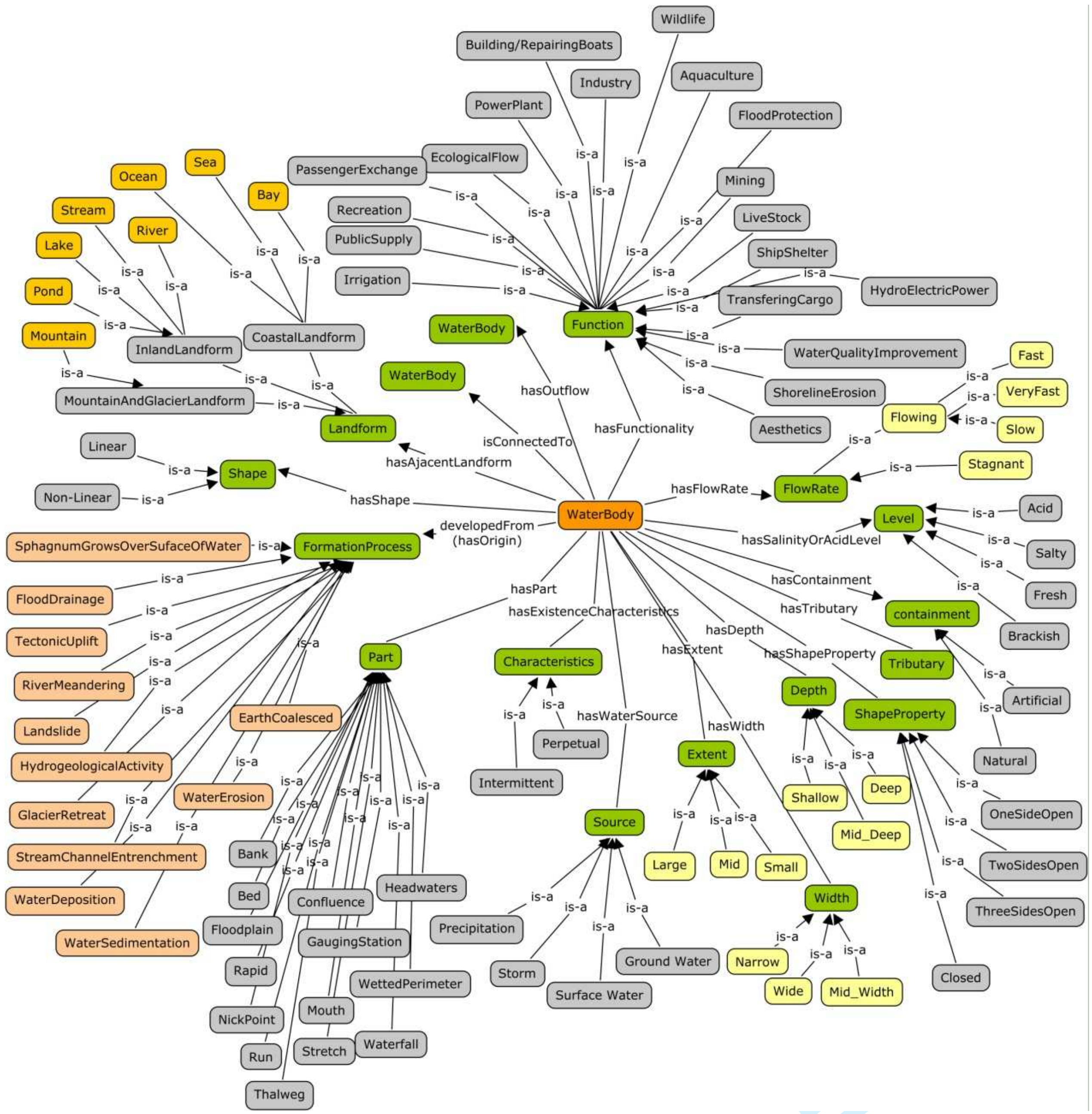


Figure 5 A Feature Space for “WaterBody”.

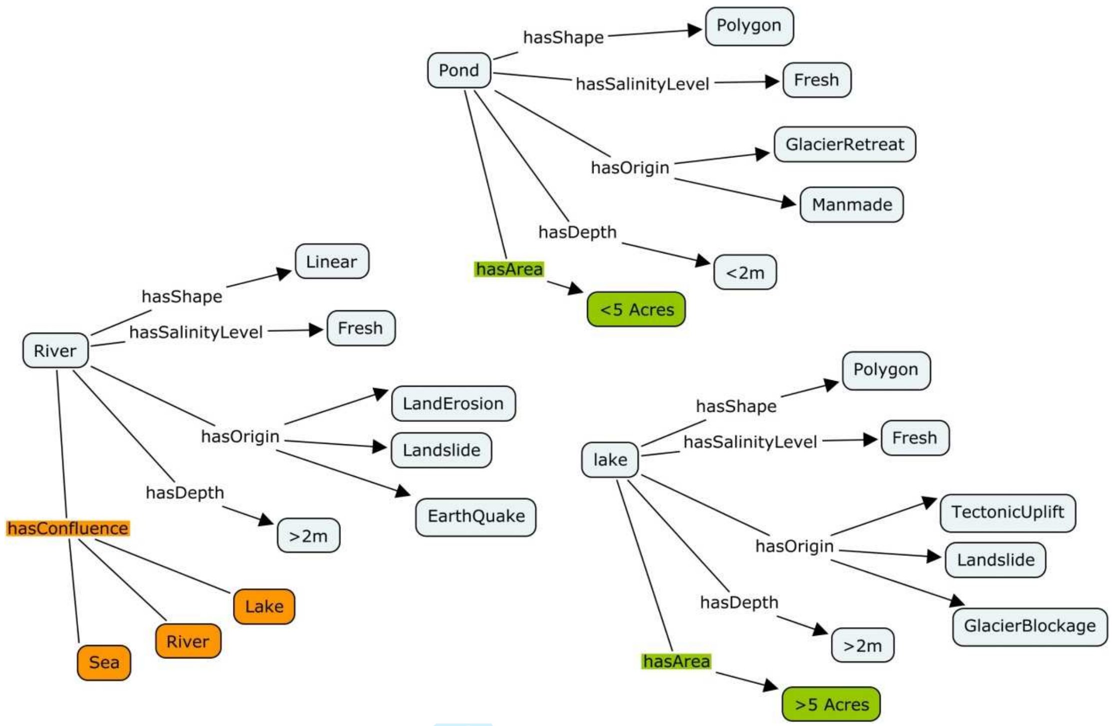


Figure 6 Semantic Definition of Three Water Body Objects.

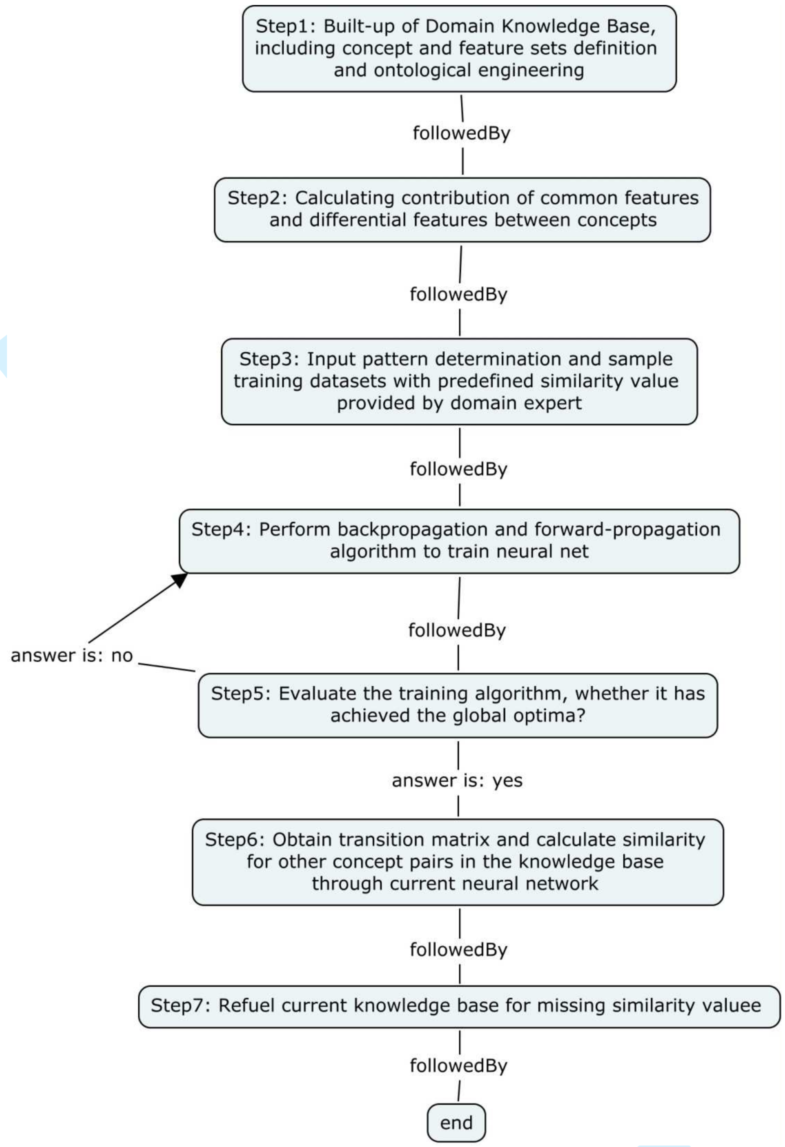


Figure 7 Training Process and Workflow.

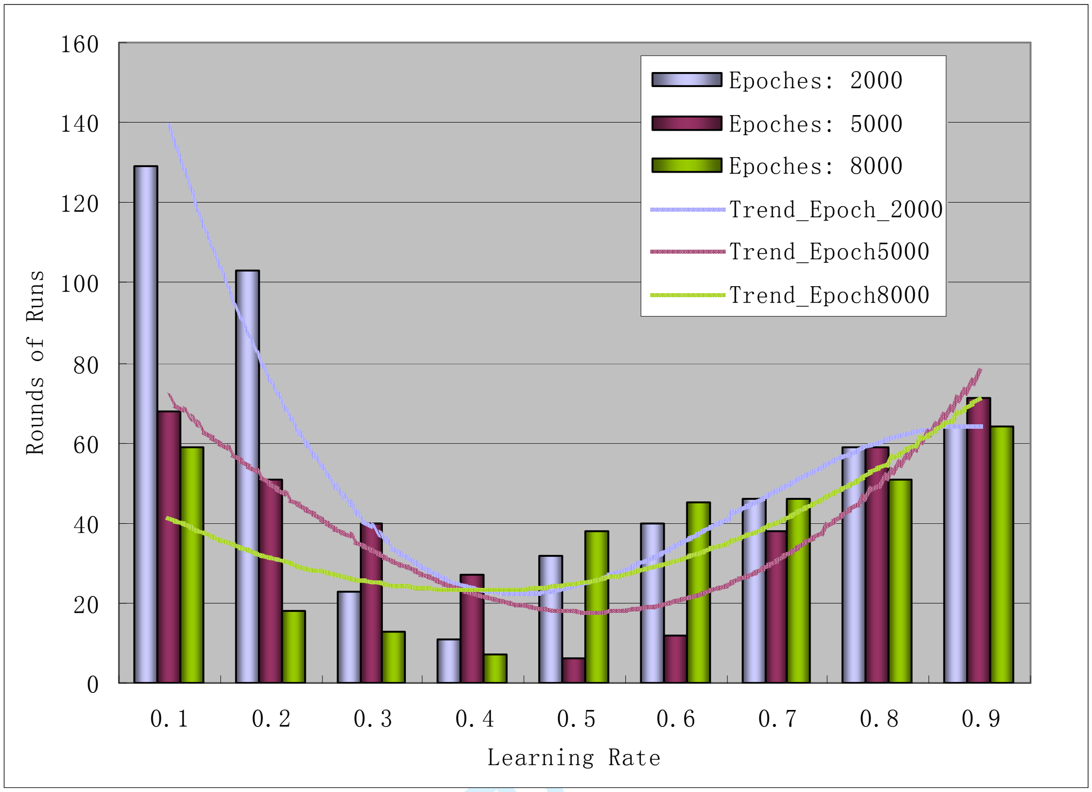


Figure 8 Number of Training Runs for the ANN Needed in Terms of Various Learning Rates.

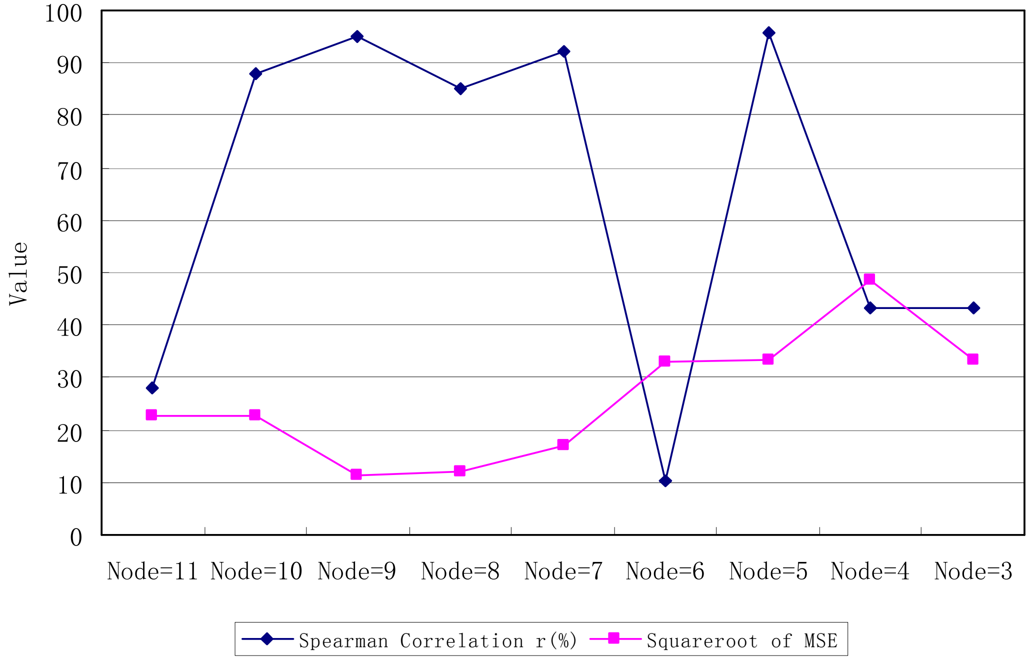


Figure 9. Prediction Accuracy as a Function of Number of Hidden Neurons.

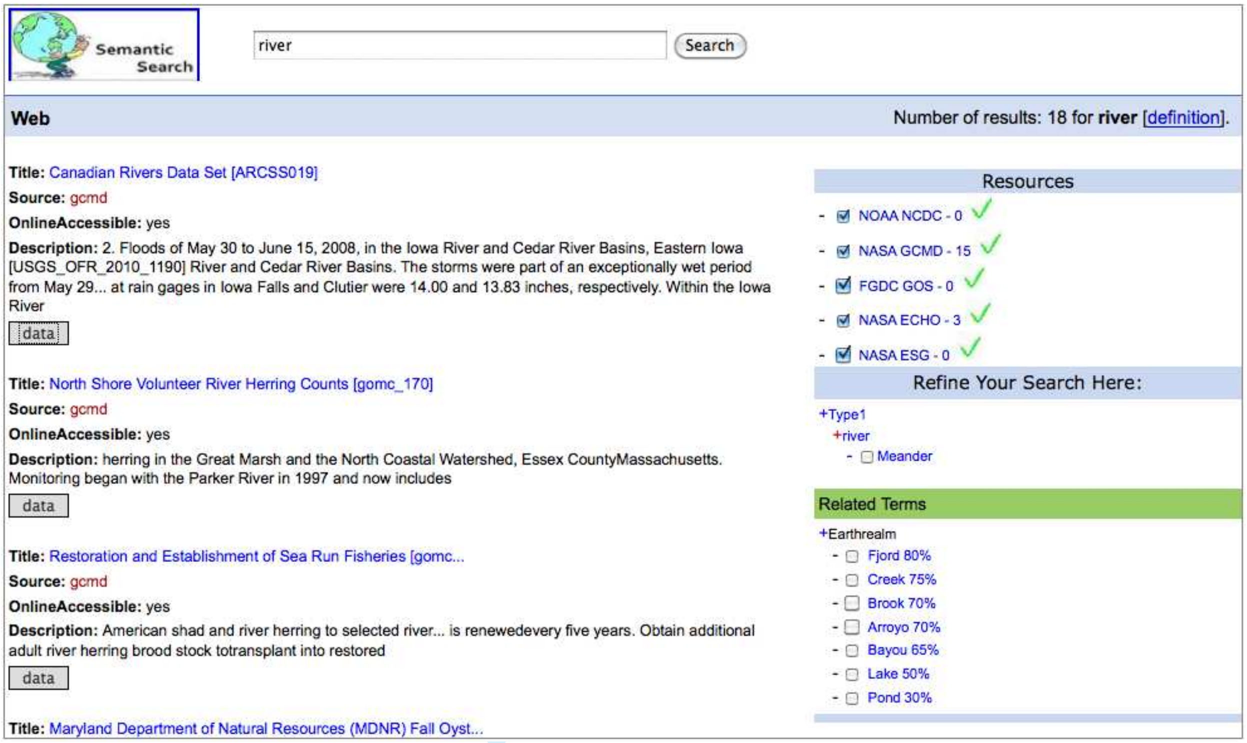


Figure 10. Screenshot of ESIP Semantic Search Testbed

List of Tables

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

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

Table 2: Case Examples for Object Pairs for Rule I and Rule II

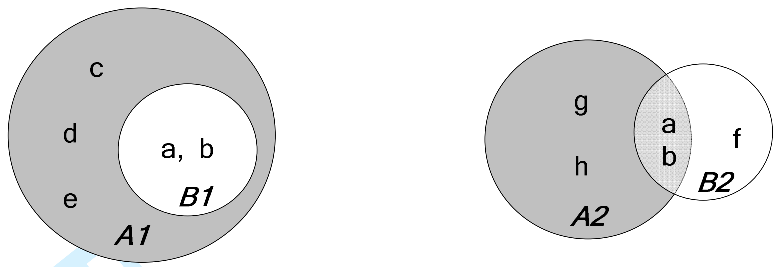
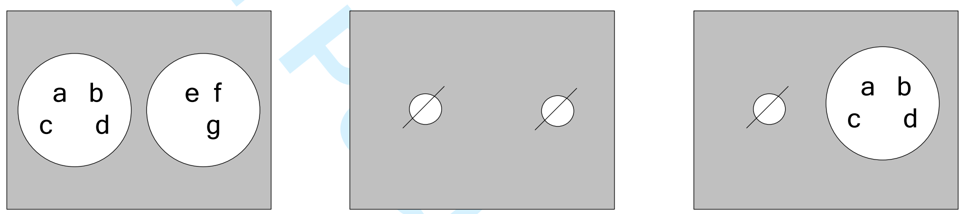
(a) (b)	Case I
(a) (b) (c)	Case II

Table 3 Survey Conducted to Human Subjects.

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

Table 4 Training parameters.

No.	Parameter	Value
1	Number of Epoches	2000 5000 8000
2	Goal of performance function	10^-3
3	Initial Learning Rate	0.1
4	Training Time	Inf.
5	momentum coefficient	0.9

Table 5 Accuracy of the ANN Prediction for Graduate Subjects and Expert Subjects.

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