# Semantic Similarity Algorithm Based on Generalized Regression Neural Network

Rui Cao[1, a], Lingda Wu[1, 2, b], Rui Wang[2, c] and Chao Yang[1, d]

[1] Equipment Academy, Beijing 101416, China

[2] The Academy of Information System and management, National University of Defense technology, Changsha 410073, China

[a]caorui_137@126.com, [b]wld@nudt.edu.cn, [c]ruiwangnudt@gmail.com, [d]yangchao@nudt.edu.cn

**Keywords:** semantic similarity, GRNN, semantic web, neural network, cross-validation.

**Abstract.** Based on the intensives study of semantic similarity algorithms and artificial neural networks knowledge, a generalized regression neural network semantic similarity algorithm is proposed. Training samples are obtained by extracting the principal component of semantic similarity influence factors; the desired spread factor and best training sample sets are gotten by cross-validation and recursive optimization; a generalized regression neural network is established with these supports. Experiment comparison and analysis verify that, the result of semantic similarity algorithm based on generalized regression neural network is more accurate than that of existing algorithms.

## Introduction

Semantic web provides great technologies to overcome the limitation of the internet, solves many interoperability application problems and produces the best possible results [1]. Machine understanding of text has acquired great interest in the research community in order to enable information extraction [2], text categorization [3], semantic annotation or anonymisation of documents [4, 5]. Semantic similarity, considered as the measure between two concepts [6], is the foundational technology of machine understanding, and many semantic similarity computational models have been proposed. Cao et al. considered the influence factors of semantic distance, information content, property, hierarchical sequence, depth information, and semantic coincidence degree between concepts, and proposed a semantic similarity algorithm based on the way of weight description [7]; Thabet summarized existing methods, concluded the semantic similarity algorithms to four categories, and evaluate them[8]; Elavarasi et al. further discussed the various available semantic similarity algorithms, and they regarded introducing more influence factors was conducive to improve the accuracy of semantic similarity algorithms [9].

At present most algorithms analyzed and described a variety of semantic similarity influence factors; they gave different weights to the influence factor description formulas, so as to get the desired results. Different from the previous research, this paper presents a semantic similarity algorithm based on generalized regression neural network (GRNN). The algorithm takes the principal component of semantic similarity influence factors as artificial neural network training samples. According to the GRNN adopted, algorithm is rational designed; cross-validation and recursive optimization are implemented, so the trained neural network owns good robustness and applicability. In this paper, a new solution is provided for semantic similarity calculation between concepts; meanwhile it promotes the exploration and application of neural network technology in the semantic web time.

## Generalized Regression Neural Network

As a variant form of Radial Basic Function neural networks, the generalized regression neural network possesses strong nonlinear approximation capability, good fault tolerance and robustness. It can achieve satisfactory results even though there is fewer sample data, and It is composed of input layer, patter layer, summation layer and output layer, as shown in Fig. 1,

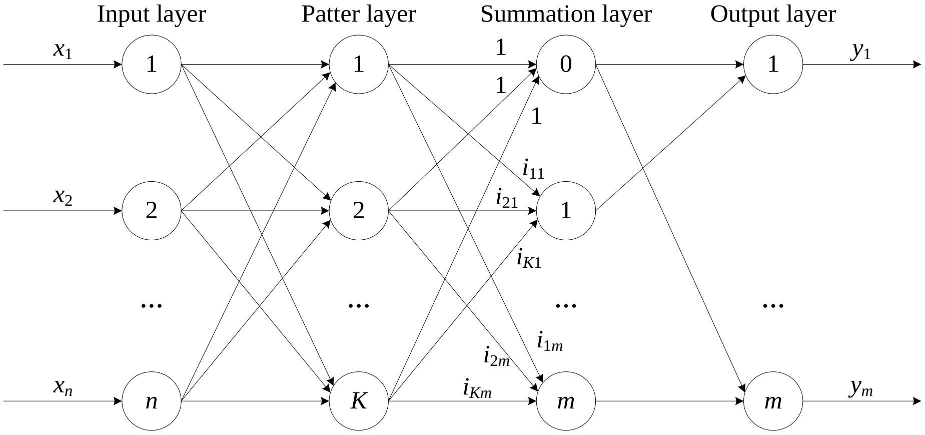
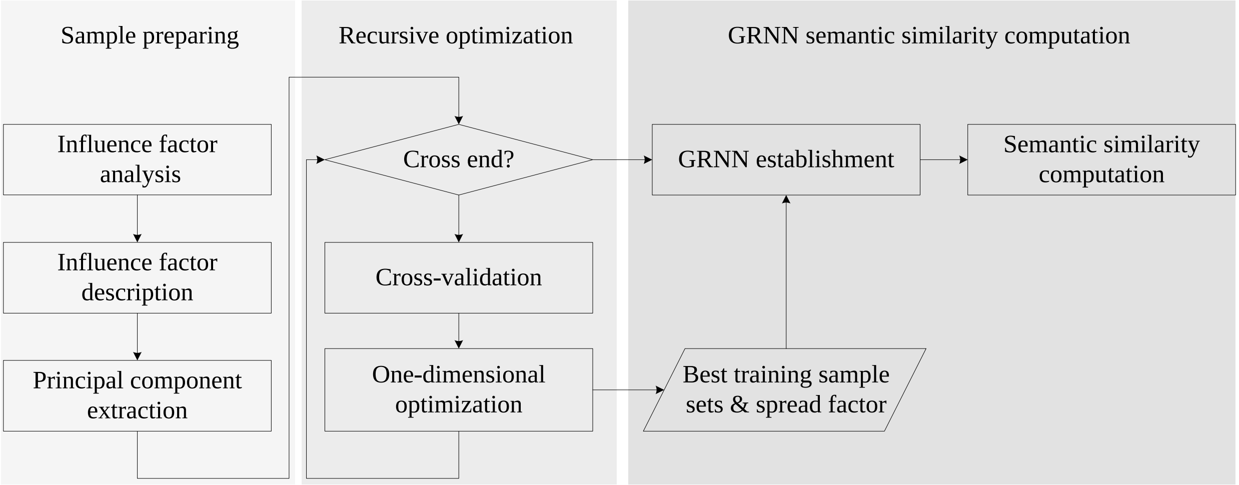


Fig. 1 Generalized regression neural network structure

## Semantic Similarity Algorithm Based on Generalized Regression Neural Network

According to the neural networks described in last section, algorithm is rationally designed to ensure that the neural network can get satisfactory results. Algorithm synthetically considers various factors affecting the semantic similarity computation; principal components extraction, cross-validation, recursive optimization and other processes are adopted to establish and train generalized regression neural network, for the semantic similarity computing between concepts. There are three main parts, and the detailed algorithm process is shown in Figure 2,



Sample preparing
Recursive optimization
GRNN semantic similarity computation
Influence factor analysis
Influence factor description
Principal component extraction
Cross end?
Cross-validation
One-dimensional optimization
GRNN establishment
Semantic similarity computation
Best training sample sets & spread factor

Fig. 2 GRNN semantic similarity algorithm process

**Sample preparing.** Selecting the correct neural network experimental sets is the premise to compute semantic similarity between concepts. Based on the influence factors analysis and description in reference [7], this algorithm mainly considers five influence factors, that is semantic distance, information content, property, hierarchical sequence and semantic coincidence degree; some factors' description formulas are improved, which make them accord with human subjective judgment better. After obtaining the neural network experimental sets by influence factors analysis and description, the algorithm extracts their principal components; it determines the number of principal components by Scree Test Criterion, and makes up the training sample sets of the principal components determined.

**Recursive optimization.** The training sample sets are divided into N repellent subsets, according to sample size. A new repellent subset is chosen as the test subset in each round of cross-validation; and the rest N-1 repellent subsets are training subsets. Spread factor is set to a fixed range of values; algorithm traverses this range with a constant step length in each round of cross-validation, and obtains the output at each spread factor. With N rounds of cross-validation, it ensures that all the repellent subsets have been the test subset. An objective function is set as the evaluation criterion of spread factor. The factor, who can achieve the optimum value of objective function, is the desired spread factor; meanwhile, the training subsets and test subset in this round of cross-validation are the best training sample sets.

**GRNN semantic similarity computation.** Based on the desired spread factor gotten from recursive optimization in last part, a general regression neural network is established; and it is trained by the best training sample sets. With regard to the semantic similarity computing between concepts, the first step is to get all the influence factor values of target concepts, by means of description formulas; and their principal components are extracted as the input data. As the neural network has already been trained by the best training sample sets, semantic similarity results can be gotten by the GRNN computation easily.

## Experimental results

The experiments were realized with Matlab and run under the experimental environment: Intel Pentium IV 3.0 GHz CPU, 1 GB RAM and Windows XP operating system.

**Experimental description.** Experimental objects were the ontology and property descriptions of Common Crops in China [10]. 42 groups of representative concepts were chosen as the experimental sets, including 32 groups training experimental sets and 10 groups testing experimental sets. On the basis of GRNN semantic similarity algorithm in last section, three or four principal components were extracted from the training experimental sets (denoted by 3PC or 4PC algorithm). The extracted principal components were identified as training sample sets, and the training sample sets were divided into four repellent subsets. The range of spread factor value was from 0.1 to 2, the step length is 0.1. With 4 rounds, different subset was chosen as the test subset of cross-validation to get the desired spread factor. On the basis of above, a generalized regression neural network was established to compute semantic similarity. Lin's algorithm [11] was chosen in the comparison experiments, as it ran compatibly and steadily in much different ontology. The Root Mean Square Error (RMSE) and Correlation coefficient were algorithm evaluation criterions. 101 independent experiments were carried out, and the experimental results as shown in the table 1,

Table 1 Semantic similarity computation results

| Number | Label | Best | Average | Worst |
|---|---|---|---|---|
| 1 | 3PC-RMSE | 0.0290 | 0.0418 | 0.0656 |
| 2 | 3PC-Correl | 0.9948 | 0.9922 | 0.9865 |
| 3 | 4PC-RMSE | 0.0308 | 0.0421 | 0.0839 |
| 4 | 4PC-Correl | 0.9952 | 0.9905 | 0.9733 |
| 5 | Lin-RMSE | — | 0.0950 | — |
| 6 | Lin-Correl | — | 0.9899 | — |

Number 1 recorded the RMSE of 50 independent experiments using semantic similarity algorithm based on GRNN; the Correlation coefficients were recorded by Number 2,and three principal components were extracted in these experiments. Number 3 and 4 respectively recorded the RMSE and Correlation coefficients of 50 independent experiments using semantic similarity algorithm based on GRNN; four principal components were extracted in these experiments. Number 5 and 6 respectively recorded the RMSE and Correlation coefficient of 1 independent experiment; Lin's algorithm was run in this experiment.

**Experimental results and analysis.** In GRNN experiments, three basic statistics of RMSE and Correlation coefficients were recorded in Table 1 Number 1~4, with the table head"best, average and worst". It was necessary to note that Lin's algorithm relied on the specific formula, so the RMSE and Correlation coefficient were determinate; As repeated exprements were unnecessary, 1 independent experiment was carried out; the RMSE and Correlation coefficient results were record in Table 1 Number 5 and 6, with the table head"averaget".

According to the experimental results, for RMSE coefficients, all the results gotten from GRNN semantic similarity algorithm proposed in this article were superior to that of Lin's; the results of 3PC algorithm were better than that of 4PC algorithm. for Correlation coefficients, the average values obtained from algorithm in this article were superior to that of Lin's; although the best values of 3PC algorithm were inferior to that of 4PC algorithm, at the aspect of average and worst vaule, 3PC algorithm is better than 4PC algorithm. On the whole, with regrad to this experimental objects, semantic similarity algorithm based on GRNN achieved better results than Lin's and it was better to extract three principal components under the same experimental conditions.

## Conclusions

This article presents semantic similarity algorithm based on generalized regression neural network. Experimental results and analysis show that, compared with the existing algorithms, this algorithm gets better results. It contributes to solving practical problems with the support of semantic similarity. More objects will be chosen as contrast experiments, and neural network should be optimized the next step

## Acknowledgements

This paper is supported by Natural Science Foundation of China (Grant 61402498 and Grant 61403404) and the National Key S&T Special Projects of "HGJ", (Grant 2013ZX01045-004-002).

## References

[1] V. Shunmughavel and P.Jaganathan. Semantic Enrichment in Ontology Mapping using Concept Similarity Computing: IEEE Fourth International Conference on Advanced Computing ( Chennai, India, December 13-15, 2012). p.1-8.

[2] D. Sánchez and D. Isern. Automatic extraction of acronym definitions from the Web: Applied Intelligence, Vol. 34 (2011) No.2. p.311-327.

[3] Q. Luo, E. Chen and H. Xiong. A semantic term weighting scheme for text categorization: Expert Systems with Applications, Vol. 38 (2011) No.10, p.12708-12716.

[4] D. Sánchez, D. Isern and M. Millán. Content annotation for the semantic web: An automatic web-based approach: Knowledge and Information Systems, Vol. 27 (2011) No. 3, p.393-418.

[5] S. Martínez, D. Sánchez, A. Valls and M. Batet. Privacy protection of textual attributes through a semantic-based masking method: Information Fusion, Vol. 13 (2012) No.4, P.304-314.

[6] D. Sánchez and M. Batet. A semantic similarity method based on information content exploiting multiple ontologies: Expert Systems with Applications, Vol.40 (2013) No.4, p.1393-1399.

[7] R. Cao and L.D. Wu. An improved semantic similarity algorithm based on domain ontology: Microelectronics &Computer, Vol. 31 (2014), No.8. p.109-114. (In Chinese).

[8] S. Thabet. Description and evaluation of semantic similarity measures approaches: International Journal of Computer Applications, Vol. 80 (2013) No. 10, p.25-33.

[9] S.A. Elavarasi, J. Akilandeswari and K. Menaga. A Survey on Semantic Similarity Measure: International Journal of Research in Advent Technology, Vol.2 (2014) No.3, p.389-398.

[10] Floras http://www.eforas.org.

[11] K. Wagh and S. Kolhe. A New Approach for Measuring Semantic Similarity in Ontology and Its Application in Information Retrieval, Lecture Notes in Computer Science, Vol.7694 (2012), pp.122-132.

[12] D. Sánchez and M. Batet. A semantic similarity method based on information content exploiting multiple ontologies, Expert Systems with Applications, Vol.40 (2013) No.4, p.1393-1399.