International Journal of Chemical and Biomolecular Science
Vol. 1, No. 3, 2015, pp. 179-184
http://www.aiscience.org/journal/ijcbs

Ontology Similarity Measuring and Ontology Mapping Algorithms Via Graph Semi-Supervised Learning

Yun Gao[1, *], Li Liang[2], Wei Gao[2]

[1]Department of Editorial, Yunnan Normal University, Kunming, China
[2]School of Information Science and Technology, Yunnan Normal University, Kunming, China

Abstract

Ontology similarity calculation is important research topics in information retrieval and widely used in biology and chemical. By analyzing the technology of semi-supervised learning, we propose the new algorithm for ontology similarity measure and ontology mapping. The ontology function is obtained by learning the ontology sample data which is consisting of labeled and unlabeled ontology data. Via the ontology semi-supervised learning, the ontology graph is mapped into a line consists of real numbers. The similarity between two concepts then can be measured by comparing the difference between their corresponding real numbers. The experiment results show that the proposed new algorithm has high accuracy and efficiency on ontology similarity calculation and ontology mapping.

Keywords

Ontology, Similarity Measure, Ontology Mapping, Semi-Supervised Learning

Received: August 3, 2015 / Accepted: August 26, 2015 / Published online: September 2, 2015

1. Introduction

As a conceptual shared and knowledge representation model, ontology has been used in knowledge management, image retrieval and information retrieval search extension. Furthermore, acted as an effective concept semantic model, ontology is employed in other fields except computer science, including medical science, social science, pharmacology science, geography science and biology science (see Przydzial et al., [1], Koehler et al., [2], Ivanovic and Budimac [3], Hristoskova et al., [4], and Kabir et al., [5] for more detail).

The ontology model is a graph $G=(V,E)$ such that each vertex v expresses a concept and each directed edge $e=v_iv_j$ denote a relationship between concepts v_i and v_j. The aim of ontology similarity measure is to get a similarity function Sim: $V\times V \rightarrow \mathbb{R}^+ \cup \{0\}$ such that each pair of vertices is mapped to a non-negative real number. Moreover, the aim of ontology mapping is to obtain the link between two or more ontologies. In more applications, the key of ontology mapping is to get a similarity function S to determine the similarity between vertices from different ontologies.

In recent years, ontology similarity-based technologies were employed in many applications. By virtue of technology for stable semantic measurement, a graph derivation representation based trick for stable semantic measurement is presented by Ma et al., [6]. Li et al., [7] determined an ontology representation method which can be used in online shopping customer knowledge with enterprise information. A creative ontology matching system is proposed by Santodomingo et al., [8] such that the complex correspondences are deduced by processing expert knowledge with external domain ontologies and in view of novel matching technologies. The main features of the food ontology and several examples of application for traceability aims is reported by Pizzuti et al., [9]. Lasierra et al., [10]

* Corresponding author
E-mail address: gaoyun@ynnu.edu.cn (Yun Gao)

pointed out that ontologies can be employed in designing an architecture for taking care of patients at home. More ontology learning algorithms can refer to [11-22].

In this paper, we present the new ontology similarity computation and ontology mapping algorithms relied on the semi-supervised learning which was proposed in Zhang et al., [23]. The basic idea of such learning approach is intending to fully use the unlabeled ontology data in the learning processing. In terms of the ontology function, the ontology graph is mapped into a real line and vertices are mapped into real numbers. Then the similarity between vertices is measured by the difference between their corresponding real numbers.

2. Main Ontology Algorithms

Let V be an instance space. For any vertex in ontology graph G, its information (including its attribute, instance, structure, name and semantic information of the concept which is corresponding to the vertex and that is contained in its vector) is denoted by a vector with p dimension. Let $v=\{v_1,\cdots,v_p\}$ be a vector which is corresponding to a vertex v. For facilitating the expression, we slightly confuse the notations and denote v both the ontology vertex and its corresponding vector. The purpose of ontology learning algorithms is to get an optimal ontology function f: $V\rightarrow\mathbb{R}^+$, then the similarity between two vertices is determined by the difference between two real numbers which they correspond to. The essence of such algorithm is dimensionality reduction, that is to say, use one dimension vector to represent p dimension vector. From this point of view, an ontology function f can be regarded as a dimensionality reduction map f: $\mathbb{R}^p\rightarrow\mathbb{R}$.

Let $K(\cdot,\cdot)$ be a positive semi-definite kernel function and K be the $n\times n$ kernel matrix with $K_{ij}=K(v_i,v_j)$. The ontology graph Laplacian matrix is defined as $L=D-K$ such that $D\in\mathbb{R}^{n\times n}$ is a diagonal degree matrix with $D_{ii}=\sum_{j=1}^{n}K_{ij}$, and the normalized version of ontology graph Laplacian is denoted by $\tilde{L}=I-D^{-1/2}KD^{-1/2}$, where I is the identity matrix. Assume that a prediction ontology function $f(\cdot)$ is evaluated on ontology data set $\{v_i\}_{i=1}^{n}$, and the prediction is expressed as $f\in\mathbb{R}^{n\times 1}$ with $f_i=f(v_i)$. The smoothness of ontology function f with regard to the ontology graph is formulated by

$$\sum_{i,j=1}^{n}\left(\frac{f_i}{\sqrt{D_{ii}}}-\frac{f_j}{\sqrt{D_{jj}}}\right)^2K_{ij}=f^T\tilde{L}f\ .$$

Its minimization is called the Laplacian regularization. For fixed labeled ontology data $\{v_i\}_{i=1}^{l}$ and unlabeled ontology data $\{v_i\}_{i=l+1}^{n}$ with u=n-l. Let $\|f\|_K$ be the Reproducing Kernel Hilbert Space norm of the prediction ontology function, γ_A be the associated balance parameter, and γ_l be the balance parameter for the smoothness. In terms of an ontology loss function $V(y,f(v))$, the Laplacian regularized semi-supervised learning in ontology setting can be denoted by $\min_f\sum_{i=1}^{l}V(y_i,f(v_i))+\gamma_A\|f\|_K^2+\gamma_l\frac{1}{n^2}f^T\tilde{L}f$ According to representer theorem, the minimizer of this optimization ontology problem has the following form:

$$f^*(v)=\sum_{i=1}^{l+u}\alpha_iK(v_i,v)$$

where α_i are called the kernel expansion coefficients.

Given an ontology data $\{v_i\}_{i=1}^{l}$ and $\{v_i\}_{i=l+1}^{n}$, then kernel matrix K, degree matrix D, the ontology graph Laplacian L and normalized graph Laplacian $\tilde{L}$ are determined. In what follows, we define $K_l=e_lK\in\mathbb{R}^{l\times n}$ as the rows in the kernel matrix corresponding to the labeled samples, where $e_l=[I_{l\times l}0_{l\times u}]$.

By virtue of representer theorem, we have $f=K\alpha$ with kernel expansion coefficient $\alpha\in\mathbb{R}^{n\times 1}$. The above ontology problem can be rewritten as $\min_{\alpha\in\mathbb{R}^{n\times 1}}\lambda_1\|K_l\alpha-y_0\|_2^2+(K\alpha)^TL(K\alpha)+\lambda_2|\alpha|_1$ with $y_0\in\mathbb{R}^{l\times 1}$ is the labels of the labeled ontology samples. Let Q= $K^TLK+\lambda_1K^Te_l^Te_lK$ and c= $K_l^Ty_0$. Then, the objective ontology function can be further denoted equivalently as

$$\min_{\alpha\in\mathbb{R}^{n\times 1}}\alpha^TQ\alpha-2c^T\alpha+\lambda_2|\alpha|_1\ . \tag{1}$$

In our article, we use low-rank approximation of symmetric, positive semi-definite matrices $K\in\mathbb{R}^{n\times n}$ such that

$$K\approx GG^T,\ G\in\mathbb{R}^{n\times m},\ m\ll n\ . \tag{2}$$

Here GG^T is the rank-m approximation of matric K. For an $n\times n$ kernel matrix, we selects a collection of m rows (or columns) $K_{nm}\in\mathbb{R}^{n\times m}$, calculate an eigenvalue decomposition on the intersection of given rows and columns $K_{mm}\in\mathbb{R}^{m\times m}$, and thus approximate the kernel matrix as

$$K\approx K_{nm}K_{mm}^{-1}K_{nm}^T\ .$$

Hence, the kernel matrix can be expressed in the following

form

$$K \approx GG^T, \quad G = K_{nm}K_{mm}^{-1/2}.$$

Let $\Theta_1 = e_l^T e_l$. The Hessian matrix (1) can be formulated by

$$\begin{aligned} Q &= K^T LK + \lambda_1 K^T e_l^T e_l K \\ &= K^T (I - D^{-1/2} K D^{-1/2} + \lambda_1 \Theta_1) K \\ &= K^T D^{-1/2} (L + \lambda_1 D \Theta_1) D^{-1/2} K \end{aligned}$$

Let $P = K^T D^{-1/2}$, $L_l = L_l + \lambda_1 D\Theta_1$. Therefore, it has a multiplicative form:

$$Q = PL_l P^T.$$

We emphasize that the degree matrix D can be approximated by

$$D \approx \mathrm{Diag}(K_{nm}K_{nm}^{-1}K_{nm}^T I). \tag{3}$$

For given l labeled ontology samples and m landmark ontology vertices, with $l<m$. Then the landmark vertex set Z are selected from two parts: (i) the labeled ontology samples; (ii) a set of unlabeled vertices by the conventional sampling plan. For deduced m land mark vertices Z, and approximate the kernel matrix K. Let $E = \tilde{D}^l(:, Z) - K_{nm}$, $W = \tilde{D}^l(Z, Z) - K_{nm}$ and

$$\tilde{D}_{ij}^l = \begin{cases} \tilde{D}_{ii}, & v_i \in Z, v_i \notin L, i = j \\ \tilde{D}_{ii}(1+\lambda_1), & v_i \in L, i = j \\ 0, & i \neq j \end{cases}. \tag{4}$$

Then both P and L_l in the matrix Q can be determined by

$$D \approx (K_{nm}K_{nm}^{-1}K_{nm}^T \tilde{D}) L_l (K_{nm}K_{nm}^{-1}K_{nm}^T \tilde{D})^T$$

$$L_l = EW^{-1}E^T. \tag{5}$$

In view of (3) and (4), the matrix can be decomposed into a low-rank factorization form as

$$D \approx FF^T$$

$$F = K_{nm}K_{nm}^{-1}K_{nm}^T \tilde{D} E W^{-1/2}. \tag{6}$$

Assume that the matrix Q in (1) is positive definite and has the eigenvalue decomposition

$$\mathrm{Q} = U \Lambda V^T. \tag{7}$$

Then the ontology problem in (1) is equivalent to the following regression:

$$\min_{\alpha \in \mathbb{R}^{n\times 1}} \|A\alpha - b\|_2^2 + \lambda_2 |\alpha|_1 \tag{8}$$

in that they have the same objective value given the same variable α. Here, ρ is a constant independent of any variables, and

$$A = \Lambda^{1/2} U^T, \quad b = \Lambda^{-1/2} U^T c. \tag{9}$$

For determined (5), the top m eigenvectors of Q can be approximated as follows. Compute the eigenvalue decomposition of the $m \times m$ matrix $FF^T = U_F \Lambda_F U_F^T$ ¼. Then the eigenvalue and eigenvectors of Q can be approximated by

$$\tilde{\Lambda} = \Lambda_F \tag{10}$$

and

$$\tilde{U} = FU_F \Lambda^{-1/2}. \tag{11}$$

The whole ontology algorithm is presented in Algorithm 1.

Algorithm 1. Given l labeled ontology samples, u unlabeled ontology samples, landmark size m, output model coefficients α

Step 1: Select the landmark set Z by using labeled ontology samples and m-l unlabeled samples.

Step 2: Determine K_{mm}, K_{nm} (2) and $\tilde{D}$ (3).

Step 3: Get low-rank form of Q by (4), (5), and (6).

Step 4: Use (10) and (11) for approximate eigenvectors of Q.

Step 5: Transform (1) by (7), (8) and (9).

Step 6: Use sparse solver to obtain optimal coefficients α. Thus get the ontology function.

Step 7: Determine the similarity between two vertices v_i and v_j according to the value of $|f(v_i) - f(v_j)|$.

3. Experiments

Two simulation experiments on ontology similarity measure and ontology mapping are designed in this section. We perform our experiments on a Windows 7 machine with 8GB RAM. And, we implement our algorithm using C++.

3.1. Ontology Similarity Measuring on Biology Data

The "Go" ontology O_1 was constructed by http: //www. geneontology. org. (Fig. 1 presents the graph structure of O_1). We use $P@N$ (defined by Craswell and Hawking [24]) to determine the equality of the experiment result.

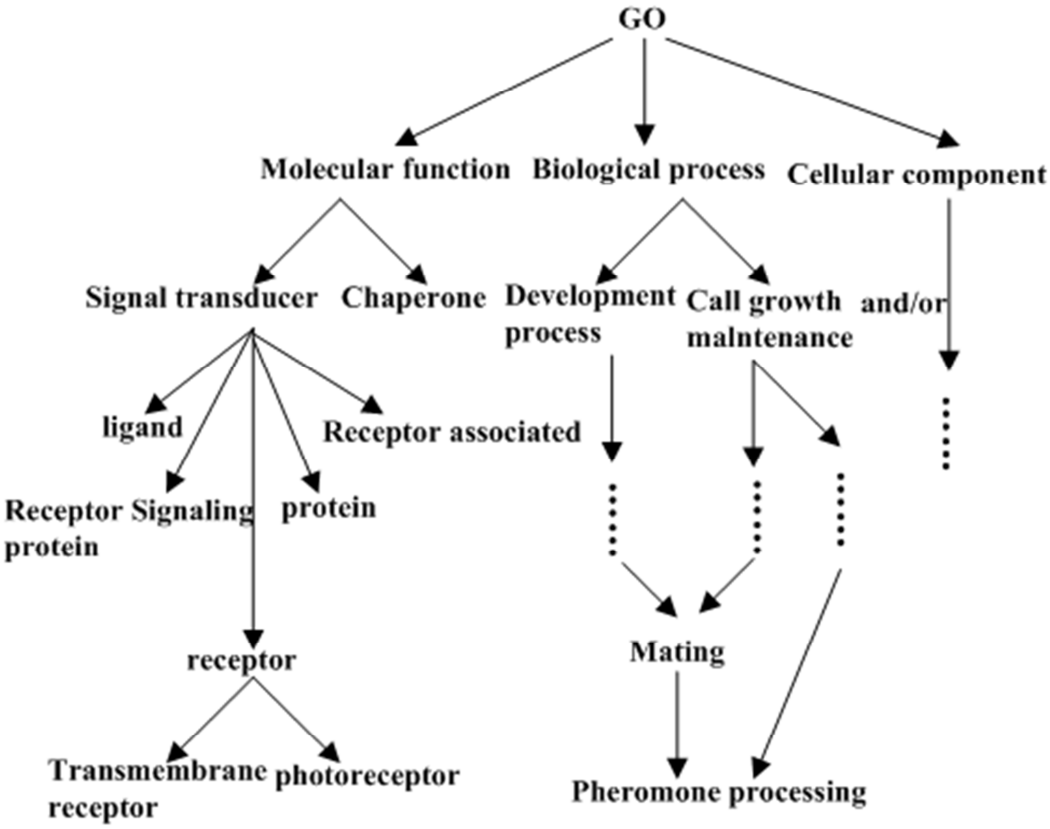

Figure 1. "Go" ontology.

The experiment shows that, *P*@3 Precision Ratio is 34.54%, *P*@5 Precision Ratio is 42.19%, *P*@10 Precision Ratio is 51.58%, *P*@20 Precision Ratio is 62.66%. Thus the algorithm have high efficient.

3.2 Ontology Mapping on Chemical Data

For our second experiment, we use "Chemical Index" ontologies O_2 and O_3 (part of the graph structures of O_2 and O_3 are presented in Fig. 2 and Fig. 3 respectively). Note that the Figure 2 and Figure 3 only show the part of vertices on O_2 and O_3. Actually, the number of vertex on O_2 is 24, and the number of vertex on O_3 is 17. The purpose of this experiment is to construct the ontology mapping between O_2 and O_3. Again, *P*@*N* criterion is applied to measure the equality of the experiment results.

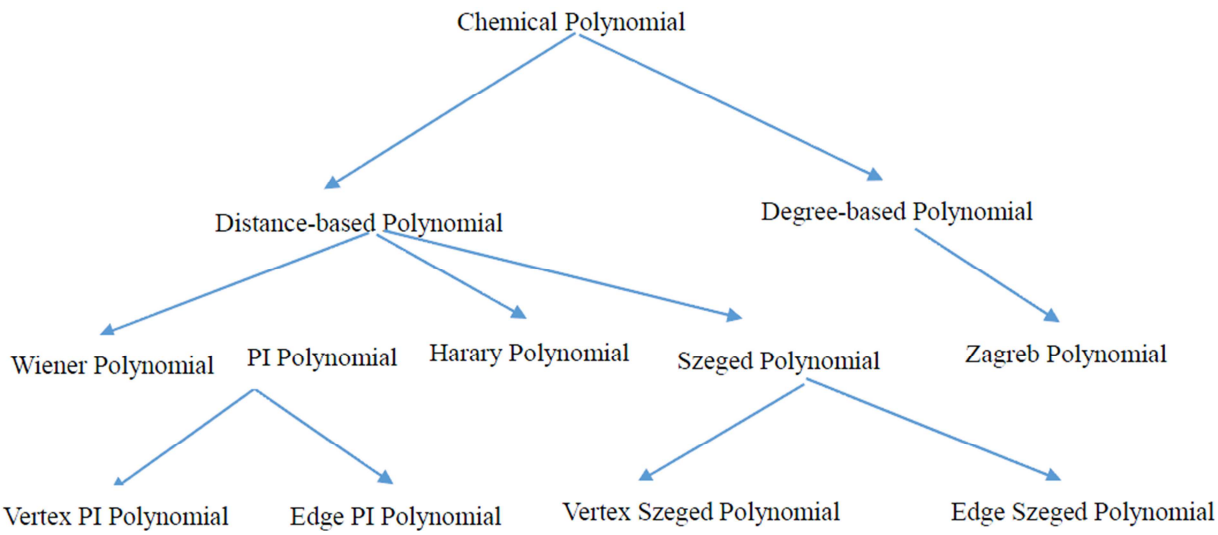

Figure 2. "Chemical Index" Ontology O_2.

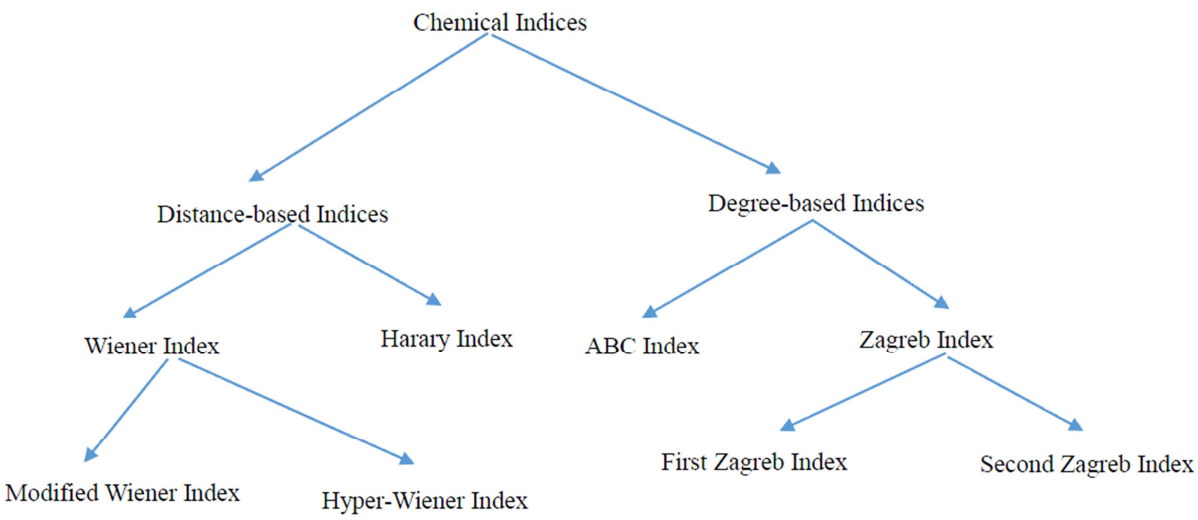

Figure 3. "Chemical Index" Ontology O_3.

The experiment shows that, *P*@1 Precision Ratio is 31.71%, *P*@3 Precision Ratio is 44.72%, *P*@5 Precision Ratio is 60.00%. Thus the algorithm have high efficient.

4. Conclusions

Ontology, as a data representation model, has been widely used in various fields and proved to have a high efficiency. The core of ontology algorithm is to get the similarity measure between vertices on ontology graph. One learning trick is mapping each vertex to a real number, and the similarity is judged by the difference between the real number which the vertices correspond to.

In our article, a new algorithm for ontology similarity measure and ontology mapping application is presented by virtue of semi-supervised learning method. Furthermore, experiment results reveal that our new algorithm has high efficiency in both biology and chemical index data. The ontology algorithm presented in our paper illustrates the promising application prospects for ontology use.

Acknowledgment

We thank the reviewers for their constructive comments in improving the quality of this paper. The research is financed by: NSFC (No.11401519).

References

[1] J. M. Przydzial, B. Bhhatarai, and A. Koleti, GPCR ontology: development and application of a *G* protein-coupled receptor pharmacology knowledge framework. *Bioinformatics*, 29(24) (2013) 3211-3219.

[2] S. Koehler, S. C. Doelken, and C. J. Mungall, The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1) (2014) 966-974.

[3] M. Ivanovic and Z. Budimac, An overview of ontologies and data resources in medical domains. *Expert Systerms and Applications*, 41(11) (2014) 5158-15166.

[4] A. Hristoskova, V. Sakkalis, and G. Zacharioudakis, Ontology-driven monitoring of patient's vital signs enabling personalized medical detection and alert. *Sensors*, 14(1) (2014) 1598-1628.

[5] M. A. Kabir, J. Han, and J. Yu, User-centric social context information management: an ontology-based approach and platform. *Personal and Ubiquitous Computing*, 18(3) (2014) 1061-1083.

[6] Y. L. Ma, L. Liu, K. Lu, B. H. Jin, and X. J. Liu, A graph derivation based approach for measuring and comparing structural semantics of ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 26(3) (2014) 1039-1052.

[7] Z. Li, H. S. Guo, Y. S. Yuan, and L. B. Sun, Ontology representation of online shopping customers knowledge in enterprise information. *Applied Mechanics and Materials*, 483(2014) 603-606.

[8] R. Santodomingo, S., Rohjans, M. Uslar, J. A. Rodriguez-Mondejar, and M.A. Sanz-Bobi, Ontology matching system for future energy smart grids. *Engineering Applications of Artificial Intelligence*, 32(2014) 242-257.

[9] T. Pizzuti, G. Mirabelli, M. A. Sanz-Bobi, and F. Gomez-Gonzalez, Food Track & Trace ontology for helping the food traceability control. *Journal of Food Engineering*, 120(1)(2014) 17-30.

[10] N. Lasierra, A. Alesanco, and J. Garcia, Designing an architecture for monitoring patients at home: Ontologies and web services for clinical and technical management integration. *IEEE Journal of Biomedical and Health Informatics*, 18(3) (2014) 896-906.

[11] Y. Y. Wang, W. Gao, Y. G. Zhang, and Y. Gao, Ontology similarity computation use ranking learning Method. *The 3rd International Conference on Computational Intelligence and Industrial Application,* Wuhan, China, 2010, pp. 20-22.

[12] X. Huang, T. W. Xu, W. Gao, and Z. Y. Jia, Ontology similarity measure and ontology mapping via fast ranking method. *International Journal of Applied Physics and Mathematics,* 1(2011) 54-59.

[13] W. Gao, and L. Liang, Ontology similarity measure by optimizing NDCG measure and application in physics education. *Future Communication, Computing, Control and Management,* 142(2011) 415-421.

[14] Y. Gao, and W. Gao, Ontology similarity measure and ontology mapping via learning optimization similarity function. *International Journal of Machine Learning and Computing,* 2(2) (2012) 107-112.

[15] X. Huang, T. W. Xu, W. Gao, and S. Gong, Ontology similarity measure and ontology mapping using half transductive ranking. *In Proceedings of 2011 4th IEEE international conference on computer science and Information technology,* Chengdu, China, 2011, pp. 571-574.

[16] W. Gao, Y. Gao, and L. Liang, Diffusion and harmonic analysis on hypergraph and application in ontology similarity measure and ontology mapping. *Journal of Chemical and Pharmaceutical Research*, 5(9) (2013) 592-598.

[17] W. Gao and L. Shi, Ontology similarity measure algorithm with operational cost and application in biology science. *BioTechnology: An Indian Journal,* 8(11) (2013) 1572-1577.

[18] W. Gao and T. W. Xu, Ontology similarity measuring and ontology mapping algorithm based on MEE criterion. *Energy Education Science and Technology Part A: Energy Science and Research*, 32(3)(2014) 3793-3806

[19] W. Gao, Y. Gao, and Y. G. Zhang, Strong and weak stability of *k*-partite ranking algorithm. *Information,* 15(11A)(2012) 4585-4590.

[20] W. Gao and T. W. Xu, Stability analysis of learning algorithms for ontology similarity computation. *Abstract and Applied Analysis,* 2013, 9 pages, http://dx.doi.org/10.1155/2013/174802.

[21] W. Gao and L. L. Zhu, Gradient learning algorithms for ontology computing. *Computational Intelligence and Neuroscience,* 2014, 12 pages, http://dx.doi.org/10.1155/2014/438291.

[22] W. Gao, L. Yan, and L. Liang, Piecewise function approximation and vertex partitioning schemes for multi-dividing ontology algorithm in AUC criterion setting (I). *International Journal of Computer Applications in Technology*, 50 (3/4) (2014) 226-231.

[23] K. Zhang, Q. J. Wang, L. Lan, Y. Sun, I. Marsic, Sparse semi-supervised learning on low-rank kernel, *Neurocomputing*, 129(2014) 265–272.

[24] N. Craswell and D. Hawking, Overview of the TREC 2003 web track. Proceeding of the Twelfth Text Retrieval Conference, Gaithersburg, Maryland, NIST Special Publication, 2003, pp. 78-92.