Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2013, 5(12):196-200

Research Article

ISSN : 0975-7384

CODEN(USA) : JCPRC5

Ontology similarity measure algorithm based on KPCA and application in biology science

[1]Xiangguang He, [2]Yaya Wang and *[3]Wei Gao

[1]Department of Experimental Training, Binzhou Polytechnic, Binzhou, China
[2]Department of Information Engineering, Binzhou Polytechnic, Binzhou, China
[3]School of Information Science and Technology, Yunnan Normal University, Kunming, China

ABSTRACT

Ontology, as a model of knowledge representation, has widely used in various fields such as chemical scienceand biology science. In this article, we present new ontology similarity calculation algorithm in terms of kernel principal component analysis and spectral cut-off regression. Then, we apply it with biology computing application.The experiment dataon "Go" ontology show the new algorithm have higher precision ratio in biology science application.

Keywords: Ontology, Similarity measure, Biology science, Go ontology, Reproducing kernel Hilbert spaces

INTRODUCTION

Ontology is a model for knowledge storing and representation, which abstracts certain application field of the real world into a set of concepts and relationships among concepts. Hence, ontology is often used in information retrieval and search expanding. By virtue of its powerful usage, ontology similarity computation has widely used in medical science, biology science and chemical science(see Lambrix and Edberg2003, Mork and Bernstein 2004,Su andGulla 2004, and Gu*et al.*, 2004 for examples). As ontology used in chemical science and biology science, every vertex can be regard as a concept of ontology, measure the similarity of vertices using the information of ontology graph[1].

Let G be an ontology graph corresponding to ontology O, the goal of ontology similarity measure is to find a similarity function Sim: $V\times V \to \mathbb{R}^{+}\cup\{0\}$which maps each pair of vertices to a real number. A populartechnology to yield optimal similarity between vertices on ontology is using a score function which maps ontology graph into a line and maps every vertex in graph into a real value. In this fashion, the similarity between vertices is transformed by measuring the difference of their corresponding real numbers[2]. Certain efficient ontology learning algorithms can refer to Wang *et al.*, 2010, Gao and Liang2011, Huang *et al.*, 2011a, Huang et al., 2011b, Gao and Lan 2011, Gao and Gao 2012. Several theoretical analyses for ontology algorithm can refer toGao and Lan 2011, Gao and Xu 2012, Gao *et al.*, 2012a, Gao *et al.*, 2012b,Gao *et al.*, 2013a,Gao *et al.*, 2013b,Gao and Xu 2013, and Yan *et al.*, 2013.

In this paper, we present a new ontology algorithm for ontology similarity measuring using the technology of kernel principal component analysis. The organization of rest paper is as follows: we describe the ontology algorithm by virtue of kernel principal component analysis; then, experiment data is given to show that our new algorithms have high accurate in biology science[3].

Algorithm for ontology similarity measure using kernel principal component analysis

We use a vector to represent the information of each vertex in ontology graph. We assume that a sample set S=(v, y)=(x_1,y_1),$\cdots$,(x_n, y_n) is selected according to an unknown distribution ρ, where $v\in V\subseteq \mathbb{R}^d$ and $y\in Y$ = [-M,M] $\subset \mathbb{R}$. The basic idea is to search a function f such that $f(v)\approx y$ and consider least squares, which can be formalized by following expected error

$$\varepsilon(f)=\int_{V\times Y}(y-f(v))^2 d\rho \text{ [4].}$$

The optimal ontology function that minimizes the expected error is the regression function $f_\rho=\int_Y yd\rho$. For given a training set S, the goal is to establish an estimator f_S whose error is close to $\varepsilon(f_\rho)$ [5].

The finding for potential solutions is usually restricted to a hypotheses space H. In this article, we consider reproducing kernel Hilbert spaces (RKHS) as hypotheses spaces. Recall that there are several well-known properties for RKHS:

- reproducing property: for $f\in H$, we have
$f(v)=\left\langle f,K(v,\cdot)\right\rangle_H$ [6];

- feature map: we can consider a mapping $\Phi: V\longrightarrow H$ which can be regard as a data parameterization related to the kernel via the following equality
$\left\langle \Phi(v_i),\Phi(v_j)\right\rangle_H=K(v_i,v_j)$, $v_i,v_j\in V$.

W. l. o. g., we assume the kernel to be continuous and bounded, i.e. $\kappa^2=\sup_{v\in V}K(v,v)<\infty$.Now, we recall the derivation of the solution to empirical risk minimization (ERM) algorithm

$$f_S=\arg\min_{f\in H}\frac{1}{n}\sum_{i=1}^{n}(f(v_i)-y_i)^2 ,$$

When H is a RKHS. Let $\Phi(v)=K(v,\cdot)=K_v$ be feature map, we infer $f(v)=\left\langle w,\Phi(v)\right\rangle$ and can easily differentiate the empirical risk with respect to w to yield a normal equation $\frac{1}{n}\sum_{i=1}^{n}\left\langle w,\Phi(v_i)\right\rangle_H \Phi(v_i)=\frac{1}{n}\sum_{i=1}^{n}y_i\Phi(v_i)$ [7].

If the data are centered, then we deduce that

$$T_v=\frac{1}{n}\sum_{i=1}^{n}\Phi(v_i)\otimes\Phi(v_i)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left\langle \cdot,\Phi(v_i)\right\rangle_H \Phi(v_i)$$

is simply the uncentered covariance operator and the solution can be denoted as w = $T_v^\dagger h_S$ with h_S = $\frac{1}{n}\sum_{i=1}^{n}y_i\Phi(v_i)$ and $T_v^\dagger$ expresses the generalized inverse of the covariance operator. Let α = $\boldsymbol{K}^\dagger\boldsymbol{y}$ and $\boldsymbol{K}^\dagger$ be the generalized inverse of the kernel matrix $[\boldsymbol{K}]_{ij}=K(v_i,v_j)$. If the Hilbert space is not finite dimensional, we can write the solution as $f(v)=\sum_{i=1}^{n}\alpha_i K(v,v_i)$ [8].

Notice that the covariance operator in the feature space under our assumptions is known to be positive and self-adjoint. Let $(\sigma_i, x_i)_{i\in I}$ be the associated eigen-system. If the data are not centered, the equivalence between principal component regression and truncated singular value decomposition cannot be assessed, unless we consider a modified kernel corresponds to following features covariance operator

$$T_v \to \hat{T}_v = (I - \frac{1_n 1_n}{n}) T_v (I - \frac{1_n 1_n}{n})$$ [9].

Since the eigenvectors of T_v and $\hat{T}_v$ can be different, spectral cut-off on the non recentered kernel is still a good algorithm but it is not clear its connection with principal component analysis.

Other thing we emphasize here is that according to the computational standpoint of view rather than working with T_v one oftendiscusses the kernel matrix since it can be presented that they share the same spectrum and their eigen-functions (eigen-vectors) are associated. For theoretical aims, it is convenient to discuss simply T_v.

The trick of kernel principal component regression can be regard as a two steps algorithm: the first step amounts to an unsupervised dimensionality reduction via kernel principal component analysis and the second step is simply ERM on the projected data. In real practice, we usually dominate the projection of the data selecting a threshold λ on the magnitude of the eigen values. More in details KPCR can be presented in the following:

Step 1: decomposition of T_v, andyield (σ_i, x_i) [10];

Step 2: Let $\vec{\varphi}^m(v) \in \mathbb{R}^m$ and $(\vec{e}_j)_j$ be a canonical basis in $\mathbb{R}^m$. Projecting the data for the first *m* components such that $\sigma_m > \lambda$ for fixed $\lambda > 0$,

$$\Phi(v) \to \vec{\varphi}^m(v) = \sum_{j=1}^{m} \langle \Phi(v), x_j \rangle \vec{e}_j$$ [11];

Step 3: Let $[\hat{\varphi}^m]_{ij} = \vec{\varphi}_j^m(v_i)$ and $[(\hat{\varphi}^m)^T \hat{\varphi}^m]_{ij} = \sigma_i \delta_{ij}$. The solution of ERM

$$\min_{\vec{w}\in\mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} (y_i - \vec{w}\cdot\vec{\varphi}^m(v_i))^2$$

is given by $\vec{w} \in \mathbb{R}^m$ can be written as

$$\vec{w} = \sum_{j=1}^{m} ([(\hat{\varphi}^m)^T \hat{\varphi}^m]^\dagger (\hat{\varphi}^m)^T \mathbf{y})_j \vec{e}_j$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{y_i}{\sigma_j} \langle \Phi(v), x_j \rangle_H \vec{e}_j$$

The solution of solution could be expressed as

$$f_S^{PCR}(v) = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{y_i}{\sigma_j} \langle \Phi(v_i), x_j \rangle_H \langle \Phi(v), x_j \rangle_H$$ [12].

For this situation, the solution $\vec{w}$ is an *m* dimensional vector.

In order to represent the spectral cut-off regularization, consider that from the formulation of ERM in the feature space, the solution on the spectrum of T_v can be rewritten as

$$w = \sum_{j=1}^{\infty}\sum_{i=1}^{n}\frac{y_i}{\sigma_j}\left\langle \Phi(v), x_j \right\rangle_H x_j \text{ [13].}$$

Hence, we derive $w^m \in H$ that $w^m = \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{y_i}{\sigma_j}\left\langle \Phi(v), x_j \right\rangle_H x_j$ .In this situation, w^m is a function in a possibly infinite dimensional space. Thus, the solution can be expressed as

$$f_S^{PCR}(v) = \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{y_i}{\sigma_j}\left\langle \Phi(v_i), x_j \right\rangle_H \left\langle \Phi(v), x_j \right\rangle_H \text{ [14].}$$

Which reveals that the solution of spectral cut-offand principal component regression are point-wise equal. In terms of RKHStheory, the obtained solutions are identical. For any $g,f\in H$ the reproducing property presents that $f(v) = g(v)\ \forall v \Leftrightarrow \left\langle f - g, K_v \right\rangle_H = 0\ \forall v$ and this fact implies that *f* and *g* are the same function[15].

EXPERIMENTAL SECTION

In this experiment, we use famous biology "Go" ontology O_1 which was constructed in http: //www. geneontology. org(Fig. 1 shows the basic structure ofO_1) for our experiment. From the experiment, we derive optimal score function for GO ontology which assign each vertex a real number. The precision ratio *P@ N* (see Craswell and Hawking 2003 for more detail) is used to measure the equality of the experiment. We first give the closest *N* concepts for every vertex on the GO ontology graph by several biology experts. Then, we determine the first *N* concepts for each vertex on ontology graph by our algorithm and deduce the precision ratio.

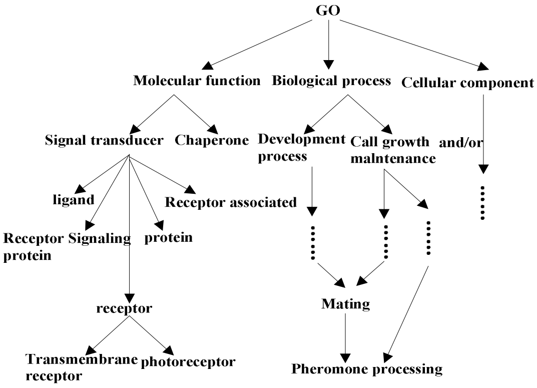

Fig. 1."Go" ontology

Simultaneously, the technologiesin Gao and Liang 2011, Gao and Gao 2012, and Huang *et al.*, 2011b are employed to the "GO" biology ontology. Calculating the accuracy by virtue of these three algorithms and comparingthe resultsto algorithm using kernel principal component analysis.Part of the data can refer to Table 1[16].

Tab. 1.The experiment results of ontology similarity measure

	P@3 average precision ratio	*P*@5 average precision ratio	*P*@10 average precision ratio	*P*@20 average precision ratio
Algorithm in our paper	52.45%	63.54%	73.56%	82.42%
Algorithm in Gao and Liang 2011	43.56%	49.38%	56.47%	71.94%
Algorithm in Gao and Gao 2012	42.13%	51.83%	60.19%	72.39%
Algorithm in Huang *et al.*, 2011b	46.38%	53.48%	62.34%	74.59%

In view of the experiment results display in Tab. 1, we arrived at the conclusion thatthe ontology similarity algorithm raised in our paper is more efficiently than algorithms presented in Gao and Liang 2011, Gao and Gao 2012, and Huang et al., 2011bespecially when *N* is lager enough. In this point of view, the new ontology algorithm for Go biology ontologyhas high efficiency.

CONCLUSION

As a data representation model, ontology has been widely used in biology science and chemical science, and proved to have a high efficiency. In this paper, we apply the trick of kernel principal component analysis to design the new ontology similarity computingalgorithm and use it in Go ontology. This new algorithm has high quality according to the experiment data above.

Acknowledgements
First, we thank the reviewers for their constructive comments in improving the quality of this paper. This work was supported in part by Binzhou Polytechnic Fund (2011XYKT01 and 2013XYKT08).We also would like to thank the anonymous referees for providing us with constructive comments and suggestions.

REFERENCES

[1] Craswell, N. and D. Hawking,.Overview of the TREC **2003** web track.In Proceeding of the Twelfth Text Retrieval Conference. Gaithersburg, Maryland, NIST Special Publication. **2003**.02.18.
[2] Gao, W., Y. Gao and L. Liang, *Journal of Chemical and Pharmaceutical Research*. **2013a**,5(9):592-598.
[3] Gao, W., Y. Gao and Y. Zhang,*Journal of Information*.**2012a**,11(A): 4585-4590.
[4] Gao, W. and M. Lan,*Microelectronics & computer*.**2011**, 28(9): 59-61.
[5] Gao,W. and L. Liang, *Future Communication, Computing, Control and Management*. **2011**,142: 415–421.
[6] Gao, W., L. Liang, T. Xu andJ. Zhou,*Journal of North University of China* (Natural Science Edition).**2013b**,34(2):140-146.
[7] Gao, W. and T. Xu, *Journal of Networks*. **2012**,8: 1251-1259.
[8] Gao,Y. and W. Gao, *International Journal of Machine Learning and Computing*. **2012**,2(2): 107-112.
[9] Gu，F., C. Cao, Y. Sui and W. Tian,*J．Comput. Sci. & Technol*. **2004**,19(2): 238-248.
[10] Huang, X., T. Xu, W. Gao and S. Gong, Ontology similarity measure and ontology mapping using half transductiveranking. In Processdings of **2011** 4th IEEE International conference on computer science and information technology. Chengdu, China,**2011.10.17**.
[11] Huang, X., T. Xu, W. Gao and Z. Jia, *International Journal of Applied Physics and Mathematics*, **2011b**, 1(1): 54-59.
[12] Lambrix, P. and A. Edberg, Evaluation of ontology tools in bioinformatics. Paci Symposium on Biocomputing, New York: IEEE Computer Society Press, **2003**.06.15.
[13] Mork, P. and P. Bernstein, Adapting a generic match algorithm to align ontologies of human anatomy. In 20th International Conf. on Data Engineering, Los Alamitos, CA, USA, Publisher: IEEE Comput. Soc. **2004**.0918.
[14] Su,X. and J. Gulla,Semantic enrichment for ontology mapping. The 9th International Conference on Information Systems (NLDB), **2004**.12.28.
[15] Wang, Y., W. Gao, Y. Zhang and Y. Gao, Ontology similarity computation use ontology learning method. In Proceedingof the 3rd International Conference on Computational Intelligence and Industrial Application, **2010**.10.12.
[16] Yan, L., W. Gao and J. Li, *Journal of Applied Sciences*, **2013**,13(16): 3257-3262.