Advanced Engineering Technology and Application
An International Journal

@ 2012 NSP
Natural Sciences Publishing Cor.

Ontology Algorithm Using Two Classes of Regressions

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

[1] *Department of Editorial, Yunnan Normal University,Kunming, Yunnan 650092 China.*
[2]*School of Information and technology, Yunnan Normal University, Kunming, Yunnan 650500, China*
Email:gaowei@ynnu. edu.cn

Received: Aug. 8, 2012; Revised Oct. 25, 2012; Accepted Dec. 30, 2012
Published online: Mar. 1, 2013

Abstract: Ontology similarity calculation is an important research topic in information retrieval and it is also widely used in computer science. In this paper, we propose new algorithms for ontology similarity measurement and ontology mapping using support vector regression and pseudo-Huber regression. Two experimental results show that the proposed new algorithm has high accuracy and efficiency on ontology similarity calculation and ontology mapping.

Keywords: ontology, ontology mapping, support vector regression, reproducing kernel Hilbert space, ε -insensitive loss, Mercer kernel, regularization parameter, Huber loss function, Pseudo-Huber loss function

1 Introduction

The problem of information retrieval is how to find the useful information for user's need from the mass of information. Information retrieval, as one of the most active branches of information theory, refers to cross-cooperation of many fields. At present, the key reason that leads the low-quality of text information retrieval is lacking of semantics retrieval tools. General speaking, information retrieval technology can be broadly divided into three categories: text retrieval, data retrieval and subject retrieval.

Text retrieval is to compare the user's query request to the form of keywords, and every word in the full text regardless the semantic matching between the query request and documents. This retrieval mode mainly based on word frequency analysis techniques. Since the mode is only aim at text matching, the searched information always large and without human intervention. The drawback is that the return of information overloads, and there are a lot of irrelevant information, users must filter from the results.

Ontology is a conceptualization clear description, it abstracts certain application field of the real world into a set of concepts and relationships of concepts. Integrating the ontology into the technology of text information retrieval not only inherit the advantages of information retrieval but also overcome the limitations that concepts information retrieval can not deal with the relationships of the concepts. It raise the accurate ratio and recall ratio of information retrieval.

As the ontology has the ability to express concept semantics through the relationship between concepts, portray the intrinsic link between concepts, and excavate those hidden and not clear concepts and information. So, it can better meet user requirements in the recall and precision aspects, and realize the retrieval intelligentize. At the same time, ontology-based retrieval methods are more in line with the of human thought, it can overcome the shortcomings of the information redundancy or information missing caused by the traditional information retrieval methods, and the query results can be more reasonable.

So, the reason for the ontology research so common. There isn't a standard communication of grammar or semantics between computer systems and people. As a model of formal shown in the conceptualization of shared, ontology provided a good way and resolved the problem to some extent. As a semantics communication method between people and machines, machinery and machines, ontology is exactly an agreement. Also, ontology is the foundation of the semantic understanding. Now, ontology similarity computation is widely used in medical science biology science [see 1] and social science [see 2]. As ontology used in information [see 3], every vertex as a concept of ontology, measure the similarity of vertices using the attraction of ontology graph. For this purpose, we do some research on the ontology-based text information retrieval based on ontology to express the text and query, to expand the specific semantic meaning of the information to be searched, to solve the problems exist in the traditional information retrieval process, to enhance the quality and efficiency of information retrieval.

Let graphs G1,G2,…,Gk corresponding to ontologies O1,O2,…, Ok, respectively, and G=G1+G2+…+Gk. For every vertex v∈V(Gi), where 1≤ i ≤k, the goal of ontology mapping is finding similarity vertices from G-Gi. So, the ontology mapping problem is also ontology similarity measure problem. Choose the parameter M∈ [0,1], let A,B are two concepts on ontology and Sim(A,B)>M, then return B as retrieval expand when search concept A. So, the ontology mapping problem is also ontology similarity measure problem. Thus, the key trick for ontology similarity measure and ontology mapping is to find the best similarity function f: V×V → $\mathbb{R}^+$ ∪ {0}, which maps each pair of vertices to a non-negative real number.

The main contribution of our paper is proposed a new ontology similarity measure method and ontology mapping using the learning method. The organization of this paper is as follows: we describe the algorithm in Section II, and two experiments are obtained in Section III which shows that the new algorithm have high quality.

2. Three classes of regression

In this paper, we study a family of ontology learning algorithms serving both purposes of support vector regression and Huber loss regression.

Loss function plays an important role in regression model. Some important loss function such as smoothed hinge loss defined as:

$$\psi_{hinge}(u)=\begin{cases} 0, & if \quad u\ge 1 \\ \frac{(1-u)^2}{2}, & if \quad 0<u<1 \\ \frac{1}{2}-u, & if \quad u\le 0 \end{cases}$$

Support vector regression is a classical kernel-based algorithm in learning theory introduced in [4]. It is a regularization scheme in a reproducing kernel Hilbert space (RKHS) HK associated with an ε -insensitive loss ψ^{ε} : $\mathbb{R} \to \mathbb{R}_+$ defined for ε ≥ 0 by

$$\psi^{\varepsilon}(u)=\max\{|u|\text{-0}\,\varepsilon,0\}=\begin{cases} |u|-\varepsilon, if\ |u|\ge\varepsilon \\ 0, \quad otherwise \end{cases} \quad (1)$$

Here, for learning functions on a compact metric space X, K : X × X → $\mathbb{R}$ is a continuous, symmetric, and positive semidefinite function called a Mercer kernel. The associated RKHS HK is defined [5] as the completion of the linear span of the set of function {Kx = K(x,·) : x∈ X} with the inner product $\langle\cdot,\cdot\rangle_K$ satisfying $\langle K_x, K_y\rangle_K = K(x,y)$. Let Y=$\mathbb{R}$ and ρ be a Borel probability measure on Z := X × Y. With a sample z= $\{(x_i,y_i)\}_{i=1}^m$ ∈ Zm independently drawn according to ρ , the support vector regression is defined as

$$f_z^{SVR}=\arg\min_{f\in H_K}\{\frac{1}{m}\sum_{i=1}^{m}\psi^{\varepsilon}(f(x_i)-y_i)+\lambda\|f\|_K^2\} \quad (2)$$

where $\lambda=\lambda(m)>0$ is a regularization parameter.

The Huber loss function describes the penalty incurred by an estimation procedure defined as [6]:

$$\psi_{\delta}(u)=\begin{cases} \frac{u^2}{2}, & if\ |u|\le\delta \\ \delta(|u|-\frac{\delta}{2}), & otherwise \end{cases} \quad (3)$$

This function is quadratic for small values of a, and linear for large values, with equal values and slopes of the different sections at the two points where $|u|=\delta$. In use, the variable u often refers to the residuals, that is to the difference between the observed and predicted values, i.e .u=y- $\hat{y}$.

The Pseudo-Huber loss function can be used as a smooth approximation of the Huber loss function, and ensures that derivatives are continuous for all degrees. It is defined as

$$\psi_{\delta}(u)=\delta^2(\sqrt{1+(\frac{u}{\delta})^2}-1) \quad (4)$$

As such, this function approximates $\frac{u^2}{2}$ for small values of u, and is parallel with slope δ for large values of u. It is applied here to a regularization scheme in the RKHS as

$$f_z^{\delta} = \arg\min_{f \in H_K} \{\frac{1}{m}\sum_{i=1}^{m} \psi_{\delta}(f(x_i) - y_i) + \lambda \| f \|_K^2\} \quad (5)$$

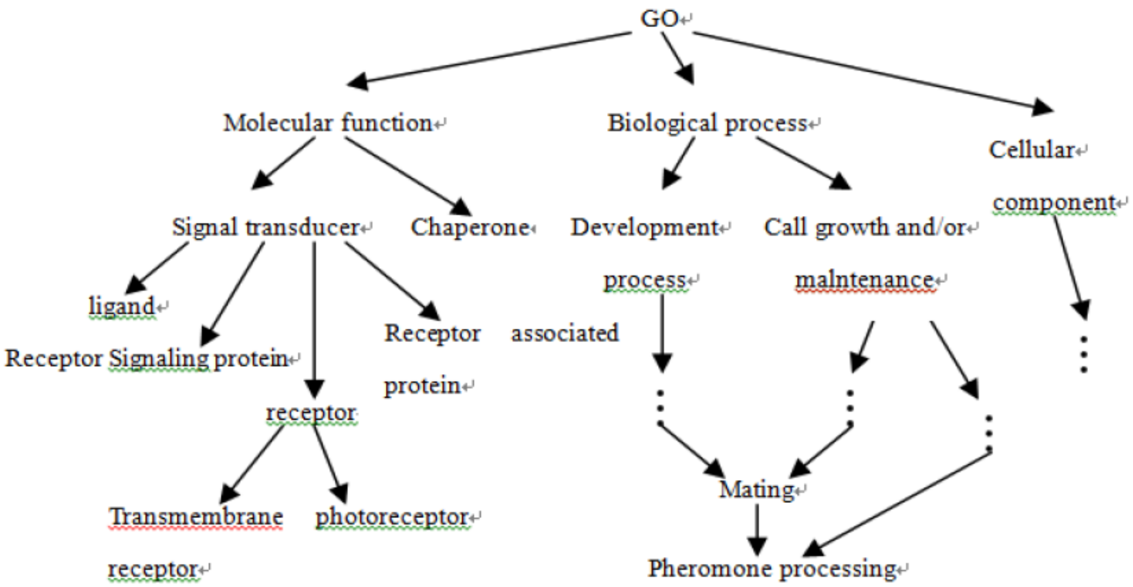

Figure 1: "GO" ontology

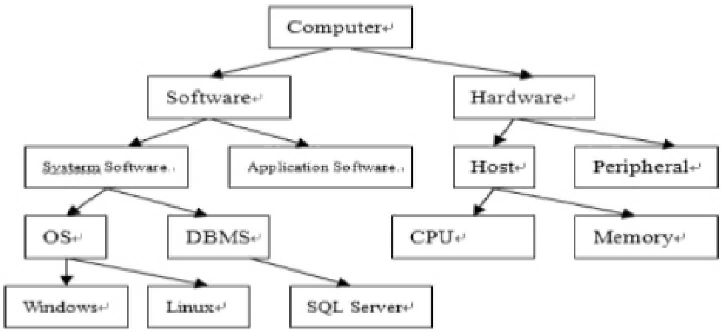

Figure 2: Computer Ontology O2

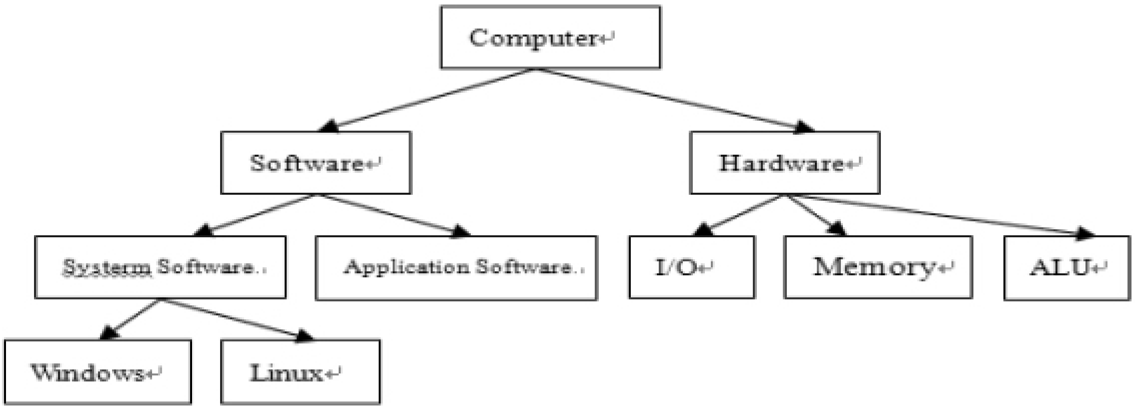

Figure 3: Computer Ontology O3

2.EXPERIMENTS

Two experiments concern ontology measure and ontology mapping are desired follow. To connect ontology to this ontology algorithm, we should use a vector to express the vertex of information. This vector contains the information of name, instance, attribute and structure of vertex, where the instance of vertex is the set of its reachable vertex in the directed ontology graph.

The first experiment concerns ontology similarity measurement is described as follows. In this experiment, we use computer ontology O1 which was constructed in [7], Fig. 1 show O1. We use P@ N (Precision Ratiosee [8]) to measure the equality of the experiment. First, the expert gives the first N concepts for every vertex on the ontology graph, and then we obtain the first N concepts for every vertex on ontology graph by the algorithm and compute the precision ratio.

Table 1: The experiment results for ontology similarity measure

	P@3 average precision ratio	P@5 average precision ratio	P@10 average precision ratio	P@20 average precision ratio
Algorithm from (2)	47.65%	54.38%	66.24%	73.83%
Algorithm from (5)	44.26%	50.39%	59.77%	68.15%

For the second experiment, we use another Computer O2 and O3, as Fig. 2 shows O2 and Fig. 3 shows O3. The goal of this experiment is given ontology mapping between O2 and O3. We also use P@ N Precision Ratio to measure the equality of experiment.

Table 2: The experiment results for ontology mapping

	P@3 average precision ratio	P@5 average precision ratio	P@10 average precision ratio	P@20 average precision ratio
Algorithm from (2)	43.26%	51.23%	59.46%	70.82%
Algorithm from (5)	39.26%	47.72%	56.72%	64.38%

Acknowledgements

First we thank the reviewers for their constructive comments in improving the quality of this paper. This work was supported in part by Key Laboratory of Educational Informatization for Nationalities, Ministry of Education, the National Natural Science Foundation of China (60903131) and Key Science and Technology Research Project of Education Ministry (210210). We also would like to thank the anonymous referees for providing us with constructive comments and suggestions.

References

[1] X. Su, and J. Gulla, “Semantic enrichment for ontology mapping”, The 9th International Conference on Information Systems (NLDB), 2004, pp. 217-228.

[2] B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt, “Ontology-based medical image annotation with description logics”, 15th IEEE International Conference on Tools with Artificial Intelligence, 2003, pp. 77-82.

[3] S. Liu , L. Chia, and S. Chan, “Ontology for naturescene image retrieval”, In on the move to meaningful Internet systems, CoopIS, DOA, and ODBASE, 2004, pp. 1050-1061.

[4] V. N. Vapnik, Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications, and Control, JohnWiley & Sons, New York, NY, USA, 1998.

[5] N. Aronszajn, “Theory of reproducing kernels,” Transactions of the American Mathematical Society, vol. 68, pp. 337–404, 1950.

[6] Huber, Peter J. (1964), "Robust Estimation of a Location Parameter", Annals of Statistics 53: 73–101

[7] http: //www. geneontology. org.

[8] N. Craswelln and D. Hawking, “Overview of the TREC 2003 web track”, Proceedings of the Twelfth Text Retrieval Conference. Gaithersburg, Maryland, NIST Special Publication, 2003, pp. 78-92.