Applied Mathematics & Information Sciences Letters
An International Journal

http://dx.doi.org/10.12785/amisl/030203

Analysis for Semi-supervised Learning Ontology Algorithm

Wei Gao[*]

School of Information, Yunnan Normal University, Kunming 650092, China

Received: 24 Jul. 2014, Revised: 8 Feb. 2015, Accepted: 10 Feb. 2015
Published online: 1 May 2015

Abstract: Ontology similarity measure and ontology mapping are widely used in knowledge representation and information processing. One method to get ontology algorithm is using graph Laplacian semi-supervised learning method, all the vertices of the ontology graph are mapped into real numbers. Then ontology similarity measure algorithm is obtained by comparing the difference of their corresponding values. In this paper, the stability of ontology algorithms is studied by adopting a strategy which adjusts the sample set by deleting one element from it. The generalized bound on such leave-one-out stability is given.

Keywords: Ontology, similarity, semi-supervised learning, graph laplacian

1 Introduction

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 intelligent. Moreover, ontology-based retrieval methods are more in line with the of human thought 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. Recently, ontology similarity computation is widely used in medical science biology science (for instance, see [1]) and social science (for instance, see [2]). As ontology used in information retrieval (for instance, see [3]), every vertex can be regard as a concept of ontology, measure the similarity of vertices using the information of ontology graph.

The key trick for ontology similarity measure is to find the best similarity function $f : V \times V \rightarrow \mathbb{R}^{+} \cup \{0\}$, which maps each pair of vertices to a non-negative real number. Gao and Liang [4] raised a ontology concept similarity method based on proximity computation. Gao, Gao and Liang [5] posed a ontology similarity measure by finding ε-neighborhood of vertices. Xu et. al. [6] gave a new ontology mapping using dimensionality reduction method. Ontology concept similarity computation algorithm with regularization framework of hypergraph was raised by [7]. And, new algorithm given by [8] using vertices matching. More details can be seen in [9], [10], [11], [12], [13], [14], [15], [16].

A ontology graph $G = (V,E)$ is a a weighted graph with $V = \{v_1, \cdots, v_n\}$ is the vertex set, E is the edge set, and a weight w_{ij} associated with edge $e_{ij} \in E$. Let $w_{ij} = 0$ if there is no edge between v_i and v_j. Assume that there is a subset of $V(G)$ whose vertices are labeled with values $y_i \in \mathbb{R}$. For all $v \in V(G)$, vector $\mathbf{v}$ represent the information of vertex v. The aim of semi-supervised learning algorithm for ontology graph is to predict the values of the rest of the vertices. In this way, all the vertices on ontology graph are mapped into real numbers, and we can get a ontology similarity measure by comparing the difference of their corresponding values.

2 Mapping Ontology Graph to the Real Line

We want to approximate a good function on a ontology graph G, with the weight matrix W_{ij}. A normal method is to consider about this function with few jumps. The standard model for seeking good functions f is by taking small values of S stated as follows

$$S(f) = \sum_{i \sim j} w_{ij}(f_i - f_j)^2.$$

[*] Corresponding author e-mail: gaowei@ynnu.edu.cn

Let $L = D - W$ be the graph Laplacian with $D = \mathrm{diag}(\sum_i w_{ii}, \cdots, \sum_i w_{ni})$. By spectral graph theory, we have

$$\sum_{i\sim j} w_{ij}(f_i - f_j)^2 = \mathbf{f}^T L\mathbf{f}.$$

Let $G = (V,E)$ be a connected ontology graph with $|V(G)| = n \neq \infty$ and the weight matrix W_{ij}. The aim is to find a function $f : V \to \mathbb{R}$. However, ontology graph is changeable model and there often new vertices add to old ontology graph. Thus, in this case, we have only partial information for old vertices. We assume the information for first k vertices are known, i.e., $f(\mathbf{v_i}) = y_i$, $1 \leq i \leq k$. The labels get form experiments on old ontology graph and can potentially with noise. Multiplicities are also allow for data vertices, i.e., each vertex of ontology graph may appear more than once with differ or same value y.

Let $\bar{y} = \frac{1}{k}\sum y_i$ and $\tilde{y} = (y_1 - \bar{y}, \cdots, y_k - \bar{y})$. There are two kinds of standard ontology algorithms:

Algorithm 1: (Ontology algorithm with parameter $\gamma \in \mathbb{R}$). Let $S = L^p$ ($p \in \mathbb{N}$) be a smoothness matrix. We add the condition $\sum f_i = 0$ for algorithm stability use. The standard model is to minimize the square loss function with smoothness penalty.

$$\tilde{\mathbf{f}} = \arg\min_{\mathbf{f}=(f_1,\cdots,f_n),\sum f_i=0} \frac{1}{k}\sum_i (f_i - \tilde{y}_i)^2 + \gamma \mathbf{f}^T L\mathbf{f}. \quad (1)$$

W.l.o.g., we always assume that the labeled vertices on ontology graph are first l ones. Since we allow vertices with different labels or the same label several times, the value of l might be distinct from k. Let $\mathbf{1} = (1,1,\cdots,1)$, $\mathbf{y} = (\sum_i y_{1i}, \sum_i y_{2i}, \cdots, \sum_i y_{mi}, 0, \cdots, 0)$, $\tilde{\mathbf{y}}$ be the n-vector, and the labels sum corresponding to the same vertex on the ontology graph. By standard linear algebra, the solution of (1) can be given as follows:

$$\tilde{\mathbf{f}} = (k\gamma S + I_k)^{-1}(\tilde{\mathbf{y}} + \mu\mathbf{1}). \quad (2)$$

Let n_i be the occurrence number of ith labeled vertex in the sample set and I_k be a diagonal matrix of multiplicities

$$I_k = \mathrm{diag}(n_1, n_2, \cdots, n_l, 0, \cdots, 0). \quad (3)$$

μ is chosen in order to $\mathbf{f} \perp \mathbf{1}$ ($\mu = 0$ means this condition is dropped). Denote linear function $s(f)$ as $s : \mathbf{f} \to \sum_i f_i$. We get $0 = s(\tilde{\mathbf{f}}) = s((k\gamma S + I_k)^{-1}\tilde{\mathbf{y}}) + s((k\gamma S + I_k)^{-1}\mathbf{1})$. Thus, we infer

$$\mu = -\frac{s((k\gamma S + I_k)^{-1}\tilde{\mathbf{y}})}{s((k\gamma S + I_k)^{-1}\mathbf{1})}.$$

Algorithm 2: (Ontology algorithm with no parameters). By assuming the values $y_1, \ldots, y_k$ with no noise, then multiple vertices in the sample set are not allowed in this case, and the ontology optimization problem is to find a smoothness function satisfying $f(\mathbf{v_i}) = \tilde{y}_i$, $1 \leq i \leq k$:

$$\tilde{\mathbf{f}} = \arg\min_{\mathbf{f}=(y_1,\cdots,\tilde{y}_k,f_{k+1},\cdots,f_n),\sum f_i=0} \mathbf{f}^T L\mathbf{f}.$$

S can be divided as

$$S = \begin{pmatrix} S_1 & S_2 \\ S_2^T & S_3 \end{pmatrix}$$

where S_1, S_2 and S_3 are $k \times k$, $k \times (n-k)$ and $(n-k) \times (n-k)$ matrix, respectively. Let $\tilde{\mathbf{f}} = (f_{k+1}, \cdots, f_n)$. Then, it follows that $\tilde{\mathbf{f}} = S_3^{-1}S_2^T((\bar{y}_1, \cdots, \bar{y}_k)^T + \mu\mathbf{1})$, and $\mu = -\frac{s(S_3^{-1}S_2^T\tilde{\mathbf{y}})}{s(S_3^{-1}S_2^T\mathbf{1})}$. Obviously, algorism 2 is the limit case of algorithm 1 when $\gamma \to 0$. The condition $\mathbf{f} \perp \mathbf{1}$ is also suggested for algorism 2 as well as algorism 1.

3 Main Results and Proof

The learning algorithm is to find a function $f_T : V \to \mathbb{R}$ for the given sample set T. To measure the quality of function, we use the generalization error $R(f)$ as follows

$$R(f) = E_\mu(f(\mathbf{v}) - y(\mathbf{v}))^2.$$

However, the underlying distribution μ on $V \times \mathbb{R}$ is unknown, and we can not compute $R(f)$ directly. Instead, we measure empirical risk $R_k(f)$ (with the square loss function) for our aim:

$$R_k(f) = \frac{1}{k}\sum_1^k (f(\mathbf{v_i}) - y_i)^2.$$

Let λ_1 be the smallest nontrivial eigenvalue of the smoothness matrix S. The main result in this paper states as follows reveal the general bound for the ontology Algorithm 1.

Theorem 3.1. Let γ be the parameter for Algorithm 1, T be a set of k vertices $v_1, \cdots, v_k$ with labels $y_1, \cdots, y_k$ which satisfy $|y_i| \leq M$, and each vertex appears no more than t times. Assuming that $\forall \mathbf{v}$, $|f_T(\mathbf{v})| \leq K$. Denote f_T as the solution of (1) using the smoothness functional S with the smallest nontrivial eigenvalue λ_1. We get with probability $1 - \delta$:

$$|R_k(f_T) - R(f_T)| \leq \beta + \sqrt{\frac{2\log(2/\delta)}{k}}(k\beta + (K+M)^2),$$

where

$$\beta = \frac{3M\sqrt{tk}}{(k\gamma\lambda_1 - t)^2} + \frac{3M}{k\gamma\lambda_1 - t}.$$

Proof. The result follows from Theorem 3.3 and Theorem 3.4 directly. □

Our result rely heavily on following definition and result.

Definition 3.2. A symmetric ontology algorithm is said to be uniformly LOO (leave-one-out) β-stable, if for any two training sets T_1 and T_2,

$$\forall \mathbf{v}, |f_{T_1}(\mathbf{v}) - f_{T_2}(\mathbf{v})| \leq \beta,$$

where T_2 is the training set such that last vertices is removed form T_1.

Theorem 3.3. (Bousquet and Elisseeff [17]) For a β-stable algorithm $T \rightarrow f_T$, we have $(\forall \varepsilon > 0)$

$$P(|R(f_T) - R_k(f_T)|$$

$$> \varepsilon + \beta) \leq 2\exp(-\frac{k\varepsilon^2}{2(k\beta + (k+M))^2}).$$

We now get the stability on ontology graph for leave-one-out case, and this result is important to get Theorem 3.1.

Theorem 3.4. (Stability on Ontology Graph for LOO Case). For sample set of size k with multiplicity of at most t, parameter γ and smoothness functional S. Assume that $k\gamma\lambda_1 - t$ is positive. Then, Algorithm 1 is a $(\frac{3M\sqrt{tk}}{(k\gamma\lambda_1 - t)^2} + \frac{3M}{k\gamma\lambda_1 - t})$-stable algorithm.

Proof. Let H be the hyperplane orthogonal to the vector $\mathbf{1}$, and P_H be the orthogonal projection on H. Then, H is invariant under S. According to (2), we have

$$(k\gamma S + I_k)\mathbf{f} = \tilde{y} + \mu\mathbf{1},$$

where μ is chosen and $\mathbf{f} \in H$. The ontology graph vertices is ordered so that the labeled vertices are in the front. Then the diagonal matrix I_k stated as (3) and $n_i \leq t$. Obviously, $l \leq k$ and the spectral radius of I_k is $\max(n_1, \cdots, n_l) \leq t$.

Moreover, the smallest nontrivial eigenvalue of S restricted to H is λ_1. According to the triangle inequality and the fact that $\|P_H(\mathbf{v})\| \leq \|\mathbf{v}\|$ for any vector $\mathbf{v}$, we infer

$$\|P_H(k\gamma S + I_k)\mathbf{f}\| \geq \|P_H k\gamma S\mathbf{f}\| - \|P_H I_k \mathbf{f}\| \geq (\lambda_1\gamma k - t)\|\mathbf{f}\|$$

holds for any $\mathbf{f} \in H$. It implies that, for restricted H, the inverse operator $(P_H(k\gamma S + I_k))^{-1}$'s spectral radius can not greater than $(\lambda_1\gamma k - t)^{-1}$.

Let $\mathbf{y}$, $\mathbf{y}'$ be the vertices vectors such that $\mathbf{y}'$ is get from $\mathbf{y}$ by removing one vertex. Thus, we denote

$$\mathbf{y} = (\sum_i y_{i1}, \sum_i y_{i2}, \cdots, \sum_i y_{il}, y_{l+1}, 0, \cdots, 0),$$

$$\mathbf{y}' = (\sum_i y_{i1}, \sum_i y_{i2}, \cdots, \sum_i y_{il}, 0, \cdots, 0).$$

The sums are taken over all values of $\mathbf{y}$ corresponding to a vertex on a ontology graph.

Let $\bar{y}$ and $\bar{y}'$ be the averages of $\mathbf{y}$ and $\mathbf{y}'$, respectively. Then, $|\bar{y} - \bar{y}'| \leq \frac{M}{k}$ and that the entries of $\tilde{y}$, $\tilde{y}'$ differ last entry, which differ by at most $M + \frac{M}{k}$. So, we obtain

$$\|\tilde{y} - \tilde{y}'\| \leq \sqrt{(M + \frac{M}{k})^2 + k(\frac{M}{k})^2} < 3M,$$

$$\mathbf{f} = (P_H(\gamma kS + I_k))^{-1}\tilde{\mathbf{y}},$$

$$\mathbf{f}' = (P_H(\gamma kS + I_k^{'}))^{-1}\tilde{\mathbf{y}}',$$

where $I_k^{'} = \mathrm{diag}(n_1, \cdots, n_{l-1}, 0, 0, \cdots, 0)$ is $n \times n$ diagonal matrices and the operator is restricted to the hyperplane H.

Let $A = P_H(\gamma kS + I_k)$, $B = P_H(\gamma kS + I_k^{'})$ restricted to the hyperplane H. With fact that $\|\|_\infty \leq \|\|$, we have

$$\mathbf{f} - \mathbf{f}' = A^{-1}\tilde{\mathbf{y}} - B^{-1}\tilde{\mathbf{y}}' = A^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{y}}') + A^{-1}\tilde{\mathbf{y}}' - B^{-1}\tilde{\mathbf{y}}'.$$

Thus,

$$\|\mathbf{f} - \mathbf{f}'\|_\infty \leq \|\mathbf{f} - \mathbf{f}'\| \leq |A^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{y}}')\| + \|A^{-1}\tilde{\mathbf{y}}' - B^{-1}\tilde{\mathbf{y}}'\|.$$

Note that the spectral radius of A^{-1} and B^{-1} are at most $\frac{1}{k\gamma\lambda_1 - t}$. We get $\|\bar{\mathbf{y}} - \bar{\mathbf{y}}'\| \leq 3M$ and

$$\|A^{-1}(\bar{\mathbf{y}} - \bar{\mathbf{y}}')\| \leq \frac{3M}{k\gamma\lambda_1 - t}.$$

Obviously, $\|\tilde{\mathbf{y}}'\| \leq 2\sqrt{kt}M$, and the spectral radius of $P_H(I_k - I_k^{'})$ smaller than 1.5, we obtain:

$$\begin{aligned}\|A^{-1}\tilde{\mathbf{y}}' - B^{-1}\tilde{\mathbf{y}}'\| &= \|B^{-1}(B - A)A^{-1}\tilde{\mathbf{y}}'\| \\ &= \|B^{-1}P_H(I_k - I_k^{'}A^{-1}\tilde{\mathbf{y}}')\| \\ &\leq \frac{3M\sqrt{tk}}{(k\gamma\lambda_1 - t)^2}.\end{aligned}$$

Combining all the fact together, we finally get

$$\|\mathbf{f} - \mathbf{f}'\|_\infty \leq \frac{3M\sqrt{tk}}{(k\gamma\lambda_1 - t)^2} + \frac{3M}{k\gamma\lambda_1 - t}.$$

□

Acknowledgment

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] P. Lambrix and A. Edberg, In: Paci.c Symposium on Biocomputing, 529-600 (2003).

[2] A. Bouzeghoub and A. Elbyed, Interoperability in Business Information Systems **1**, 73-84 (2006).

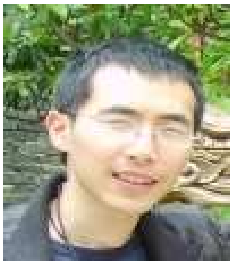
[3] X. Su and J. Gulla, In Proc.The 9th International Conference to Information Systems 217-228 (2004).

[4] W. Gao and L. Liang, Journal of Changchun University, **19**, 12-14 (2009).

[5] W. Gao, Y. Gao, and L. Liang, Journal of Yunnan Normal University (Natural Science Edition) **31**, 37-40 (2011).

[6] T. Xu, J. Gan, L. Gao, and W. Gao, Journal of Northwest Normal University (Natural Science Edition) **47**, 52-55 (2011).

[7] W. Gao and L. Liang, Microelectronics and Computer **28**, 15-17 (2011).
[8] W. Gao and L. Liang, Journal of Anhui University (Natural Science Edition) **34**, 28-31 (2010).
[9] W. Gao and M. Lan, Microelectronics and Computer **28**, 59-61 (2011).
[10] X. Huang, T. Xu, W. Gao, and Zhiyang Jia, International Journal of Applied Physics and Mathematics **1**, 54-59 (2011).
[11] W. Gao and L. Liang, Future Communication, Computing, Control and Management **142** 415-421 (2011).
[12] Y. Gao and W. Gao, International Journal of Machine Learning and Computing **2**, 107-112 (2012).
[13] X. Huang, T. Xu, W. Gao, and Shu Gong, Journal of Engineering **1**, 20-24 (2012).
[14] W. Gao, L. Zhu, and L. Liang, Journal of Southwest University (Natural science edition) **31**, 118-121 (2012).
[15] W. Gao, L. Zhu, and Y. Zhang, Journal of Southwest China Normal University (Natural science edition) **34**, 64-67 (2011).
[16] Y. Wang, W. Gao, Y. Zhang, and Y. Gao, Intelligent computation and industrial application, June, Hong Kong, China, Publisher: International Industrial Electronic Center. **III**, 20-23 (2011).
[17] O. Bousquet and A. Elissee, Advances in Neural Information Processing Systems **13**, 196-202 (2001).



Wei Gao, male, was born in the city of Shaoxing, Zhejiang Province, China on Feb.13, 1981. He got two bachelor degrees on computer science from Zhejiang industrial university in 2004 and mathematics education from College of Zhejiang education in 2006. Then, he was enrolled in department of computer science and information technology, Yunnan normal university, and got Master degree there in 2009. In 2012, he got PHD degree in department of Mathematics, Soochow University, China. Now, he acts as lecturer in the department of information, Yunnan Normal University. As a researcher in computer science and mathematics, his interests are covering two disciplines: Graph theory, Statistical learning theory, Information retrieval, and Artificial Intelligence.