昆明学院学报　2018，40（3）：65～69　ISSN 1674－5639　CN 53－1211/G4
Journal of Kunming University

非线性框架下本体稀疏向量学习算法

龚　澍[1]，邹目权[2]，高　炜[3*]
（1. 广东科技学院 计算机系，广东 东莞 523083；2. 昆明学院 信息技术学院，云南 昆明 650214；
3. 云南师范大学 信息学院，云南 昆明 650500）

摘要：本体作为一种高效的语义模型，被广泛应用于工程科学的各个领域，而语义相似度计算是本体算法的核心内容. 利用本体稀疏向量得到本体相似度计算的策略可用于高维数据和大数据处理. 因此，考虑在非线性框架下的本体稀疏向量计算算法，用平方亏损函数表示误差项，通过近端梯度的计算得到对应的迭代策略. 最后，通过两个实验来说明该本体稀疏向量学习算法对于特定的工程应用中本体相似度计算和本体映射是有效的.

关键词：本体；相似度计算；本体映射；稀疏向量；非线性
中图分类号：TP393.092　**文献标识码**：A　**文章编号**：1674－5639（2018）03－0065－05
DOI：10.14091/j.cnki.kmxyxb.2018.03.012

Ontology Sparse Vector Learning Algorithm in Nonlinear Setting

GONG Shu[1]，ZOU Muquan[2]，GAO Wei[3*]
（1. Department of Computer Science，Guangdong University Science and Technology，Dongguan，Guangdong，China　523083；
2. College of Information Technology，Kunming University，Kunming，Yunnan，China　650214；
3. College of Information，Yunnan Normal University，Kunming，Yunnan，China　650500）

Abstract：As an efficient semantic model，ontology is widely used in various fields of engineering science，and the core of the ontology algorithm is semantic similarity calculation. The ontology similarity calculation via sparse vector is a strategy which can be used for high-dimensional data and big data processing. So the ontology sparse vector calculation algorithm nonlinear framework is considered，and the corresponding iteration strategy is obtained by using squared loss function to express the error term and calculating the proximal gradient. Lastly，two experiments are described to show the efficiency of ontology sparse vector learning algorithm for ontology similarity computation and ontology mapping in specific engineering applications.

Key words：ontology；similarity measure；ontology mapping；sparse vector；nonlinear

在当代大数据存储、管理和计算中，为了更好地表示数据之间的关联，要求对应数据模型具有结构化存储数据的特征，并且这种数据模型要求易于计算、统计和分析. 因此，作为一种结构化概念共享、存储模型，本体越来越受到数据信息管理者的重视，并逐渐成为近年来数据信息领域研究的热点问题. 一般地，本体概念的结构化存储方式可以用层次图来表示，即可用一个图来表示一个本体，其中图中的每个顶点概念对应一个概念或者一条信息，顶点之间的边代表概念或者信息之间的某种隐含的关系，比如从属关系等.

除了结构化存储之外，数据模型要求有利于学者对得到的数据信息进行计算、统计、推理并最后得到一些结论. 因此，在本体上的算法一般围绕着本体中存储的数据信息展开，即找出它们的相互关系. 从这一角度来说，本体概念之间（即本体图上顶点之间）的相似度计算成为本体工程应用的核心算法. 而在大数据背景下，一个本体图往往存

收稿日期：2017－07－06
基金项目：国家自然科学青年基金资助项目（11401519）.
作者简介：龚澍（1985—），女，江西吉安人，讲师，硕士，主要从事机器学习、人工智能研究.
[*] **通讯作者**：高炜（1981—），男，浙江绍兴人，副教授，博士，主要从事统计学习理论研究，E-mail：gaowei@ynnu.edu.cn.

储了海量的信息（比如生物基因 GO 本体和植物学 PO 本体），这导致传统启发式地设计本体概念计算公式的方法已经无法胜任大数据处理的本体框架. 因此通过机器学习得到本体概念相似度计算方法已成为近年来本体算法研究的主流.

近年来，一些学者研究了如何通过学习方法得到本体算法. 例如：从矩阵论的角度出发，提出本体中相似度矩阵学习的策略[1]；应用成对排序学习方法得到对应的本体相似度计算和本体映射算法[2]；将原有基于正则化模型的本体学习算法加以改进，得到在有噪声条件下同样适用的新本体学习算法[3]；从另一个角度对基于正则化模型的本体学习算法加以改进，使其适用于半监督学习框架[4]；基于 BMRM 迭代排序方法的本体学习算法[5]；将 Mahalanobis 矩阵学习融入到本体算法中，得到对应的本体相似度计算和本体映射算法[6]；在多重分割的框架下提出新的本体优化框架[7]；在多重分割的框架下得到基于无穷推进策略的本体学习算法[8]；用核函数作为相似度计算函数，通过学习方法得到最有相似度核，进而得到本体相似度计算函数[9]；得到基于 ADAL 方法的本体稀疏向量学习算法，并通过稀疏向量来计算顶点之间的相似度[10].

而本文考虑在一种特殊框架下的本体稀疏向量学习，即通过本体稀疏向量得到对应实数值的计算公式中存在非线性函数，我们称其为非线性框架. 将利用梯度计算方法得到该框架下的一种迭代策略，并利用实验验证算法的有效性.

1 基于稀疏向量的本体算法框架

为了使本体模型适应学习算法框架的要求，首先需要对概念信息进行数值化处理，即对每个顶点而言，用一个向量来表示这个顶点的所有信息. 为了方便起见，约定用 v 同时表示顶点和它对应的向量，因此下文中的 v 在不引起混淆的情况下同时可以理解为一个顶点和它对应的向量，不再用标准向量的粗体表示. 在本体概念向量表示下，概念之间的相似度通过向量之间的几何距离来衡量，距离越小则相似度越大，距离越大则相似度越小.

设 $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_d)^T\in\mathbb{R}^d$ 为本体稀疏向量，特别地 $\boldsymbol{\beta}^*=(\beta_1^*,\cdots,\beta_d^*)^T\in\mathbb{R}^d$ 为最优本体稀疏向量. 本体稀疏向量的特点是绝大多数分量的值为 0. 对于本体中的顶点 $v=(v_1,\cdots,v_d)$，通过本体稀疏向量得到其对应实数值的计算方法如下：

$$y=v^T\boldsymbol{\beta}^*+\varepsilon, \tag{1}$$

其中 ε 是代表偏移量或者误差量的值. 基于稀疏向量的本体算法，其基本思想是通过样本学习得到最优本体稀疏向量 $\boldsymbol{\beta}^*$，再由（1）式，通过本体图中每个顶点计算出它们对应的实数，然后本体概念间的相似度就可通过它们对应顶点的对应实数之间的一维距离来判定：距离越小，相似度越大；距离越大，相似度越小. 该算法的核心思想也是一种降维思想，首先将每个本体顶点对应的语义信息用一个 d 维向量来表示，通过本体稀疏向量进行降维度，将原来 d 维向量转化为一维实数 y. 从这一层意义上说，本体稀疏向量在计算模型中是作了桥梁的作用，且 $\boldsymbol{\beta}^*$ 的优劣直接影响到相似度计算结果. 因此，基于本体稀疏向量的本体算法，其核心是本体稀疏向量的学习.

近年来，出于本体在生物学、医学等领域的广泛应用，本体所涉及的表示信息量越来越大，因此对本体算法的要求也越来越高. 基于本体稀疏向量的本体算法，其优势在于在某个具体应用中，可以对该种应用无关的特征信息进行有效屏蔽，而突出有价值的信息. 比如在遗传学中，某种遗传病只与若干个基因有关，与大部分基因无关，而稀疏向量则能有效地寻找目标基因，达到预期的目标. 由于稀疏算法对高维数据降维有着神奇的效果，而对于低维数据，稀疏算法反而会增加算法的复杂度. 因此，本文中可设向量的维度远远大于本体样本的容量，即 $d>>n$.

本文考虑一种（1）式的变化模型如下：

$$y=f(v^T\boldsymbol{\beta}^*)+\varepsilon, \tag{2}$$

其中 f 可理解为一类转换函数，当 f 为恒等函数时，（2）式即退化为（1）式. 因而，计算模型（2）式可以理解为（1）式的一种扩展. 设 $\{v_i,y_i\}_{i=1}^n$ 为服从某种独立同分布的本体样本集，最优本体稀疏向量可通过以下正则化模型得到：

$$\boldsymbol{\beta}^*=\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^n(y_1-f(v_i^T\boldsymbol{\beta}))^2+\lambda\|\boldsymbol{\beta}\|_1, \tag{3}$$

其中 λ 为平衡参数，$\|\boldsymbol{\beta}\|_1$ 为控制本体稀疏向量的稀疏度，$\lambda\|\boldsymbol{\beta}\|_1$ 称为平衡项，而作为主体部分，$\frac{1}{n}\sum_{i=1}^n(y_1-f(v_i^T\boldsymbol{\beta}))^2$ 为误差项. （3）式的本质是使用平方亏损作为亏损函数的误差表示.

2 新算法描述

本文将讨论非线性表示下（f 为非线性函数），最优本体稀疏向量的求解. 以下均假设给定的本体样本 $\{v_i,y_i\}_{i=1}^n$ 满足 $y_i=f(v_i^T\boldsymbol{\beta}^*)+\varepsilon_i$，函数 f 为单调连续可导. 设平方亏损部分为：

$$L(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}(y_i-f(v_i^T\boldsymbol{\beta}))^2. \tag{4}$$

事先假设 $\boldsymbol{\beta}^*$ 是稀疏的，并通过（3）式来估计 $\boldsymbol{\beta}^*$ 的值.

由于 f 为非线性函数，$L(\boldsymbol{\beta})$ 有可能是非凸的. 需找到一个驻点（又称为平衡点）$\hat{\boldsymbol{\beta}}$ 满足 $\lambda\boldsymbol{\xi}+\nabla L(\hat{\boldsymbol{\beta}})=0$，其中 $\nabla L(\hat{\boldsymbol{\beta}})$ 表示 $L(\hat{\boldsymbol{\beta}})$ 的梯度且 $\boldsymbol{\xi}\in\partial\|\hat{\boldsymbol{\beta}}\|_1$. 下面利用近端梯度（Proximal gradient）方法来得到这个驻点. 该方法可以得到一个迭代序列 $\{\boldsymbol{\beta}^{(t)},\ t\geqslant 0\}$，其中

$$\boldsymbol{\beta}^{(t+1)}=\underset{\beta\in\mathbb{R}^d}{\operatorname{argmin}}\{<\nabla L(\beta^{(t)}),\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)}>+\frac{\alpha_t}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1, \tag{5}$$

在（5）式中，$\|\cdot\|_2$ 表示标准欧几里得泛数，$\alpha_t>0$，$1/\alpha_t$ 表示 t 次迭代的次步长，而 $\nabla L(\boldsymbol{\beta}^{(t)})$ 的值可以通过如下方法计算：

$$\nabla L(\boldsymbol{\beta}^{(t)})=-\frac{1}{n}\sum_{i=1}^{n}(y_i-f(v_i^T\boldsymbol{\beta}^{(t)}))f'(v_i^T\boldsymbol{\beta}^{(t)})v_i.$$

记　　$\boldsymbol{u}^{(t)}=\boldsymbol{\beta}^{(t)}-\frac{1}{\alpha_t}\nabla L(\boldsymbol{\beta}^{(t)})$，

则（5）式的解可表示为：

$$\boldsymbol{\beta}_i^{(t+1)}=S(u_i^{(t)},\frac{\lambda}{\alpha_t}), \tag{6}$$

其中 $1\leqslant i\leqslant d$，$S(\cdot,\cdot)$ 称为软边界算子，定义为 $S(u,a)=\operatorname{sign}(u)\max\{|u|-a,0\}$.

下面给出本文基于近端梯度计算的本体稀疏向量学习算法：

输入　平衡参数 $\lambda>0$，更新因子 $\eta>0$，参数 $\zeta>0$，$\alpha_{\min}$ 和 $\alpha_{\max}$ 满足 $0<\alpha_{\min}<1<\alpha_{\max}$，整数 $M>0$，函数 $\varphi(\boldsymbol{\beta})=L(\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_1$；

初始化　$t\leftarrow 0$ 并选择 $\boldsymbol{\beta}^{(0)}\in\mathbb{R}^d$；

步骤 1　使用如下方法计算步长 α_t 的值. 输入迭代计数值 t，$\delta^{(t)}=\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}$ 和 $\boldsymbol{g}^{(t)}=\nabla L(\boldsymbol{\beta}^{(t)})-\nabla L(\boldsymbol{\beta}^{(t-1)})$. 若 $t=0$，则 $\alpha_t=1$；否则 $\alpha_t=\frac{\langle\delta^{(t)},\boldsymbol{g}^{(t)}\rangle}{\langle\delta^{(t)},\delta^{(t)}\rangle}$ 或 $\alpha_t=\frac{\langle\boldsymbol{g}^{(t)},\boldsymbol{g}^{(t)}\rangle}{\langle\delta^{(t)},\boldsymbol{g}^{(t)}\rangle}$；

步骤 2　重复更新 $\boldsymbol{u}^{(t)}\leftarrow\boldsymbol{\beta}^{(t)}+\frac{1}{n\alpha_t}\sum_{i=1}^{n}(y_i-f(v_i^T\boldsymbol{\beta}^{(t)}))f'(v_i^T\boldsymbol{\beta}^{(t)})v_i$，$\boldsymbol{\beta}_i^{(t+1)}\leftarrow S(u_i^{(t)},\frac{\lambda}{\alpha_t})$，$\alpha_t\leftarrow\eta\alpha_t$. 期间若发现 $\boldsymbol{\beta}^{(t+1)}$ 满足 $\varphi(\boldsymbol{\beta}^{(t+1)})\leqslant\max\{\varphi(\boldsymbol{\beta}^{(j)})-\zeta\frac{\alpha_t}{2}\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\|_2^2;\max(t-M,0)\leqslant j\leqslant t\}$，则结束更新；

步骤 3　更新迭代计数值 $t\leftarrow t+1$；

步骤 4　若 $\frac{\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}\|_2}{\|\boldsymbol{\beta}^{(t)}\|_2}$ 足够小，则输出 $\hat{\boldsymbol{\beta}}\leftarrow\boldsymbol{\beta}^{(t)}$，否则返回到步骤 1.

通过上述迭代算法得到最优本体稀疏向量 $\boldsymbol{\beta}^*$ 的近似解，并通过（2）式得到每个顶点对应的 y 值. 设 y_i 和 y_j 分别是本体顶点 v_i 和 v_j 对应的实数，则 v_i 和 v_j 对应本体概念之间的相似度通过 $|y_i-y_j|$ 的值来衡量，其值越小，相似度越大；其值越大，相似度越小.

3　实验

下面两个实验将分别验证新本体学习算法对本体相似度计算和构建本体映射的效率.

3.1　本体相似度计算实验

本文所采用的数据是来自于 http://www.plantontology.org 网站构建的植物学 PO 本体 O_1（其基本结构可参考图 1）. 该本体可以看成是一个植物学数据库，我们用来检验本文新算法对相似度计算的效率. 为了对算法的效率有一个比较，将以下 3 类经典本体学习算法也作用于 PO 本体：基于一般排序学习方法的本体算法[11]、基于快速排序学习的本体算法[12]和基于 NDCG 测度计算的本体算法[13]. 全部实验结果使用 $P@N$[14] 平均准确率来评判. 将 3 类经典本体学习算法得到的 $P@N$ 准确率与本文新本体算法得到的 $P@N$ 准确率进行对比，取 $N=3$，5，10 时的对比，数据如表 1 所示.

表 1　实验 1 部分数据

算法名称	$P@3$ 平均准确率/%	$P@5$ 平均准确率/%	$P@10$ 平均准确率/%
本文算法	49.52	66.99	89.67
一般排序算法	45.49	51.17	58.59
快速排序算法	42.82	48.49	56.32
NDCG 本体算法	48.31	56.35	68.71

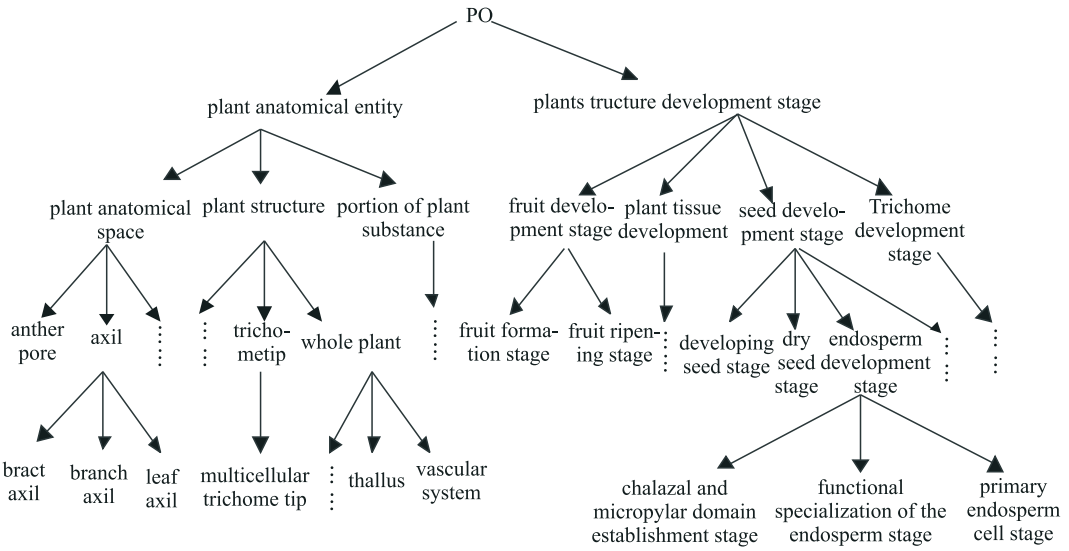

图1 PO本体O_1

通过上述表 1 中取 $N=3$，5，10 时 $P@N$ 准确率对比可知，本文所提出的心本体学习算法对于植物学 PO 本体上进行相似度计算而言，随着 N 的增大，其效率明显高于其他 3 种经典本体学习算法.

3.2 本体映射实验

接下来，使用下面两个“仿生机器人”本体 O_2 和 O_3 来验证本文新本体学习算法对构建基于相似度计算的本体映射（即对于某个顶点而言，在另一个本体中找到与其相似度相对较大的顶点，作为映射结果返回给用户）的效率（图 2 和图 3）. 为了让数据有所对比，我们还将基于 k－部排序的本体学习算法[15]、基于 NDCG 测度计算的本体学习算法[13]和基于超图调和分析的本体学习算法[16]也作用于“仿生机器人”本体 O_2 和 O_3，之后将这 3 类学习算法得到的 $P@N$ 准确率与本文新本体学习算法得到的 $P@N$ 准确率进行对比. 取 $N=1$，3，5 时的数据比较，如表 2 所示.

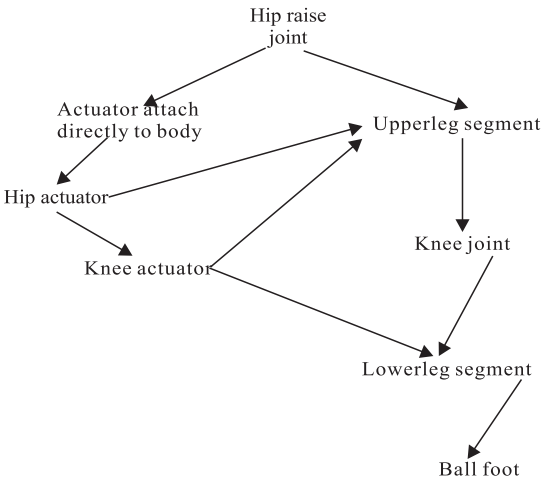

图2 “仿生机器人”本体O_2

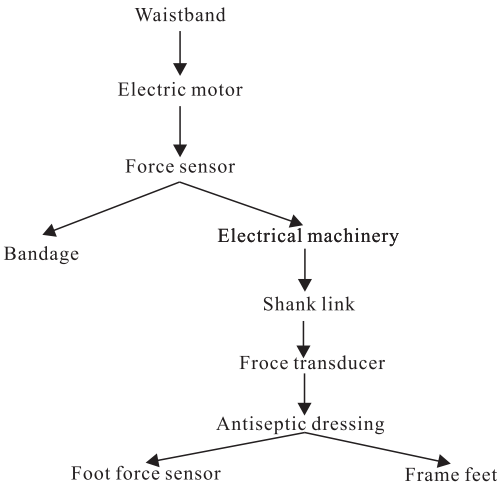

图3 “仿生机器人”本体O_3

表 2 实验 2 部分数据

算法名称	$P@1$ 平均准确率/%	$P@3$ 平均准确率/%	$P@5$ 平均准确率/%
本文算法	27.78	53.70	80.00
k-部排序本体算法	27.78	48.15	54.44
NDCG 本体算法	22.22	40.74	48.89
调和分析本体算法	27.78	46.30	53.33

根据上表 2 在 $N=1$，3，5 时的 $P@N$ 准确率数据对比可以看出，随着 N 的增大，本文新本体学习算法对于在“仿生机器人”本体 O_2 和 O_3 间建立基于相似度的本体映射的效率要高于另外 3 类算法.

4　结语

本体是一种集数据结构化存储、分析、计算、统计、推理于一体的数据管理模型. 在实际工程应用中，本体算法的核心是相似度计算. 由于其具备强大的数据管理和应用功能，目前本体早已应用于制药学、社会科学、GIS、管理学等. 本文利用近端梯度计算方法，得到非线性框架下最优本体稀疏向量的迭代求解算法，并因此得到相似度计算策略. 最后两个实验数据充分说明，新本体学习方法对于在植物学领域中进行相似度计算和在仿生机器人领域两个本体之间构建基于相似度的本体映射，都有较高的效率.

[参考文献]

[1] 吴剑章，朱林立，高炜. 本体算法中相似度矩阵的学习[J]. 小型微型计算机系统，2015，36（4）：773－777.

[2] 朱林立，戴国洪，高炜. 成对排序本体学习算法[J]. 西南师范大学学报（自然科学版），2013，38（12）：101－106.

[3] 朱林立，吴访升，叶飞跃，等. 有噪条件下基于正则化模型的本体学习算法[J]. 西北师范大学学报（自然科学版），2014，50（6）：41－45.

[4] 朱林立，戴国洪，高炜. 正则化框架下半监督本体算法[J]. 微电子学与计算机，2014，31（3）：126－129.

[5] 朱林立，高炜. 基于 BMRM 迭代排序方法的本体学习算法[J]. 科学技术与工程，2013，13（13）：3653－3657.

[6] 吴剑章，余晓，高炜. 基于 Mahalanobis 矩阵的学习的本体算法[J]. 西南大学学报（自然科学版），2015，37（2）：117－122.

[7] ZHU L L，GAO W. Ontology similarity measuring and ontology mapping based on new optimization model in multi-dividing setting [J]. Journal of Computational Information Systems，2015，11（1）：377－386.

[8] GAO W，ZHU L L，GUO Y. Multi-dividing infinite push ontology algorithm [J]. Engineering Letters，2015，23（3）：132－139.

[9] ZHU L L，MIN X Z，GAO W. Algorithm of ontology similarity measure based on similarity kernel learning [J]. International Journal of Computers & Technology，2015，14（12）：6304－6309.

[10] GAO W，ZHU L L，WANG K Y. Ontology sparse vector learning algorithm for ontology similarity measuring and ontology mapping via ADAL technology [J]. International Journal of Bifurcation and Chaos，2015，25（14）：383－392.

[11] WANG Y，GAO W，ZHANG Y，et al. Ontology similarity computation use ranking learning method [C] // The 3rd International Conference on Computational Intelligence and Industrial Application，2010：20－22.

[12] HUANG X，XU T，GAO W，et al. Ontology similarity measure and ontology mapping via fast ranking method [J]. International Journal of Applied Physics and Mathematics，2011，1（1）：54－59.

[13] GAO W，LIANG L. Ontology similarity measure by optimizing NDCG measure and application in physics education [J]. Future Communication，Computing，Control and Management，2011，142：415－421.

[14] CRASWELL N，HAWKING D. Overview of the TREC 2003 web track [C] //Proceedings of the 12th Text Retrieval Conference. Maryland：NIST Special Publication，2003：78－92.

[15] 兰美辉，任友俊，徐坚，等. k-部排序本体相似度计算[J]. 计算机应用，2012，32（4）：1094－1096.

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