

基于超图正则化模型的本体概念相似度计算

高 炜, 梁 立

(云南师范大学 计算机科学与技术学院, 云南 昆明 650092)

摘 要: 概念的语义相似度研究, 是知识表示以及信息检索领域中的一个重要内容. 选择核函数计算本体图边的权值. 求解超图正则化模型得到优化函数, 从而将本体图中每个顶点映射成一个实数, 通过比较实数间的差值判断两概念的相似程度. 实验表明该方法对于计算本体概念间的相对相似度是有效的.

关键词: 本体; 正则化模型; 高斯核; 逆多元二次核

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2011)05-0015-03

Ontology Concept Similarity Computation Based on Regularization Framework of Hypergraph

GAO Wei, LIANG Li

(School of Computer Science and Information Technology, Yunnan Normal University, Kunming 650092, China)

Abstract: Study semantic similarity of the concept is an important aspect in knowledge representation and information retrieval. Compute the edge weights using kernel function. Obtain the optimization function using hypergraph regularization framework. The function maps the vertices of graph to real numbers. Calculate the relative similarities of concepts by comparing the difference of corresponding values. Experiments show that the method for calculating the relative similarity between the concepts of ontology is efficient.

Key words: ontology; regularization framework; gaussian kernel; inverse multiquadric kernel

1 引言

本体(Ontology)中两个概念 A, B 的相似度记为 $\text{Sim}(A, B) [0, 1]$. 根据特定的应用背景设定阈值 M , 当 $\text{Sim}(A, B) > M$, 认为 A, B 之间有较高的相似度. 在信息检索中, 如果查找 A , 则可将 A 进行查询扩展, 将所有满足 $\text{Sim}(A, B) > M$ 的 B 返回给用户. 因此本体概念相似度的计算好坏直接影响用本体进行信息检索的效率.

传统的本体概念相似度计算算法存在以下问题: (1) 需要人为设定的参数较多; (2) 需要两两计算相似度, 计算量大且不够直观. 为此, 文中提出基于超图正则化模型的本体相似度计算算法, 其思想是: 首先对本体层次有向图及代表概念的顶点进行预处理, 用基础无向图代表原来的有向图并向量化

顶点的信息; 其次, 根据不同的应用背景选择适当的核函数计算每条边的权值; 再次, 通过求解超图正则化模型得到最优函数 f , 同时将本体图映射成实直线; 最后, 通过比较顶点对应实数间的差值得到概念间的相对相似度.

2 超图正则化模型

设 $G = (V, E)$ 为超图, E 是超边集合, $e \in E$ 为 V 的一个子集. 带权超图是指给每条超边 e 赋一个正值 $w(e)$. 顶点 v 与超边 e 相关联若 $v \in e$. $\forall v \in V$, v 的度定义为 $d(v) = \sum_{\{e \in E | v \in e\}} w(e)$. 对于超边 e , 它的度定义为所包含的顶点个数, 记为 $\delta(e) = |e|$. 实值函数 $f: V \rightarrow R$ 可视为 $R^{|V|}$ 上的向量. 它将超图中每一个顶点映射成一个实数, $f(v)$ 的值即为向量中对

收稿日期: 2010-06-18; 修回日期: 2010-09-01

基金项目: 国家自然科学基金项目(60903131)

应顶点 v 的分量. 对于分类问题, 可假定 f 的每个分量的值为 1 或 -1; 对于回归问题, $f(v)$ 的值为实数. 超图学习的目标就是要找出最优的 f , 为此文献 [1] 提出的超图正则化模型如下:

$$\arg \min_f \Omega(f) = \frac{1}{2} \sum_{e \in E} \frac{1}{\delta(e)} \sum_{\{u, v\} \subseteq e} w(e) \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2$$

它的含义是: 对于分类问题, 要使同类顶点之间的边有高权重, 异类顶点之间的边有低权重; 对于回归问题, 使同一条超边上的点对应值尽可能接近.

超图 G 的 $|V| \times |E|$ 关联矩阵 H 定义为: $h(v, e) = \begin{cases} 1, & v \in e \\ 0, & v \notin e \end{cases}$. 对角矩阵 D_v, D_e 对角线上的元素分别为顶点的度和边的度. 设 W 为权值对角矩阵. G 的邻接矩阵 $A = HWH^T - D_v$, 其中 H^T 表示 H 的转置. 定义矩阵 $\Delta = I - D_v^{-1/2} H W D_e^{-1} H^T D_v^{-1/2}$, 这里 I 表示单位矩阵. 从而有^[1]

$$\Omega(f) = \langle f, \Delta f \rangle = f^T \Delta f \quad (1)$$

由此可知, Δ 是一个半正定对称矩阵. Δ 的最小特征值为 0, 对应的特征向量为 $\sqrt{d}$.

超图正则化模型在诸多领域有高效的应用, 相关内容参考文献 [1-4]. 与之相关的研究参考文献 [5-7]. 文中将运用此模型计算本体概念的相似度.

3 算法描述

通过上述分析可知, 函数 f 将超图的每个顶点映射成一个实数, 同时使图邻接关系越近的顶点对应的实数差值越小. 这和本体层次结构图有反向相似之处, 本体图中邻接关系越近的顶点所对应概念的语义相似度往往越大. 受文献 [1] 启发, 文中给出基于超图正则化模型的本体概念相似度计算算法.

3.1 本体层次有向图预处理和边权值计算

首先将本体层次有向图去掉每条边的方向得到基础图 G , 并将其替换原来的图. 再将每个顶点对应概念的所有信息用一个向量表示. 选择适当的核函数计算每一条无向边的权值. 设 $uv \in V(G)$, 可使用以下几种核函数来计算 $w(uv)$:

$$\text{高斯核 } w(uv) = e^{-\frac{\|u-v\|^2}{\sigma^2}} \quad (2)$$

逆多元二次核:

$$w(uv) = (c^2 + \|u-v\|^2)^{-\alpha} \quad (3)$$

式中, c 和 α 均为正数.

3.2 正则化模型的求解

由于经过预处理后的本体图 G 为无向简单图,

因此在超图正则化模型中取 $\delta(e) = 2$ 可得对应无向简单图的正则化模型如下:

$$\arg \min_f \Omega'(f) = \frac{1}{4} \sum_{uv \in E} w(uv) \left(\frac{f(u)}{\sqrt{d(u)}} - \frac{f(v)}{\sqrt{d(v)}} \right)^2 \quad (4)$$

注意到此时 $D_e = 2I$, 从而 $\Delta = I - \frac{1}{2} D_v^{-1/2} H W H^T D_v^{-1/2} = I - \frac{1}{2} D_v^{-1/2} (D_v + A) D_v^{-1/2} = \frac{1}{2} (I - D_v^{-1/2} A D_v^{-1/2})$. 以每条边的权值为基础, 将本体概念间的相似度计算问题转化为求概念顶点所对应实数的差值问题. 即求解正则化模型 (4), 找到最优 f 将每个顶点映射成一个实数, 两概念对应实数的差值越小, 它们的相似度就越大. 问题归结于:

$$\arg \min_{f \in \mathbb{R}^{|V|}} \Omega'(f) \quad (5)$$

Subject to $|f| = 1, \langle f, \sqrt{d} \rangle \geq 0$

其中限制 $\|f\| = 1$ 是为了删除成比例的 f , 限制 $\langle f, \sqrt{d} \rangle \geq 0$ 是为了得到非平凡的 f , 即避免 f 为特征值 0 对应的特征向量. 由式 (1) 可知在限制条件下最小化 $\Omega'(f)$ 的问题可等价求解 Δ 的次小特征值对应的特征向量 (即需找 Δ 的不为 0 的最小特征值对应的特征向量, Δ 也称为拉普拉斯矩阵).

3.3 相似概念的选择

得到 f 后, 对于 $v \in V(G)$, $f(v)$ 的值即为向量 f 对应于该顶点的分量. 于是最优函数 f 将本体图 G 映射成实直线, 将图中顶点映射成对应的实数. 用实数间的差值来衡量顶点所代表概念间的相似程度. 对于顶点 v , 可用下面的方法选择与其相似的顶点, 并将对应的概念返回给用户作为查询扩展的结果:

方法 1: 选择阈值 M , 返回集合 $\{u \in V(G), |f(u) - f(v)| \leq M\}$.

方法 2: 选择自然数 N , 返回与 v 最相似的 N 个顶点所对应的概念.

可见, 方法 1 兼顾公平, 方法 2 可以控制返回概念的数量. 方法的选择及 M, N 的选取由具体的应用背景而定. 下面我们通过具体的实验来验证上述算法对于特定的应用领域是高效的.

4 实验分析

本试验构建了如下“数学学科”本体, 如图 1 所示. 预处理后用高斯核函数计算每条边的权值. 由于本算法是通过比较两个概念对应实数的差来判断两个概念是否相似, 因此本算法得到的是相对相似度而非绝对相似度. 由此我们采用 $P@N^{[8]}$ 准确率对

实验结果进行评价,即首先由领域专家主观判断某概念在本体中最相似的前 N 个概念,然后由算法计算出与该概念最相似的 N 个概念,看是否与领域专家的评判一致,计算准确率。

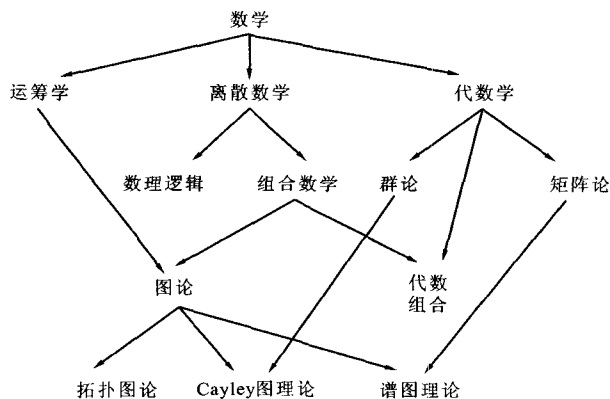


图1 “数学学科”本体结构图

实验结果显示, $P@1$ 平均准确率达 61.5%, $P@3$ 平均准确率达 76.9%, $P@5$ 平均准确率达 86.2%。从而该算法是有效的。

5 结束语

文中给出了基于超图正则化模型的本体概念相对相似度计算算法。本算的优点在于:(1)由于算法将每个顶点映射成了实数,因此不需要类似传统算法的成对计算相似度,降低了计算量;(2)算法充分利用了本体自身的结构,将图映射为实直线,更加直观地表现了各概念间的近似程度;(3)与传统本体相似度计算方法相比,本算法需要领域专家确定的参数大大减少。

参考文献:

- [1] Zhou D, Huang J, Schölkopf B. Beyond pairwise classification and clustering using hypergraphs[P]. Canada: University of Waterloo, 2005.
- [2] Zhou Dengyong, Huang Jiayuan, Bernhard Schölkopf. Learning with Hypergraphs: Clustering, Classification, and Embedding[C]// Proceedings of 20th Annual Conference on Neural Information Processing Systems, 2006, Vancouver / Whistler, Canada; IEEE, 2006:1601—1608.
- [3] Liang Sun, Shuiwang Ji, Jieping Ye. Hypergraph spectral learning for multi-label classification[C]// Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA; ACM, 2008:668—676.
- [4] Chen Gang, Zhang Jianwen, Wang Fei, et al. Efficient multi-label classification with hypergraph regularization[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA; IEEE, 2009 (9):1658—1665.
- [5] Sameer Agarwal, Kristin Branson, Serge Belongie. Higher order learning with graphs[C]// Proceedings of the 23rd International conference on Machine Learning, Pittsburgh, PA; ICML, 2006: 17—24.
- [6] Zhou Dengyong, Huang Jiayuan, Bernhard Schölkopf. Learning from Labeled and Unlabeled Data on a Directed Graph[C]// Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany; ICML, 2005:1041—1048.
- [7] Isaac Pesenson. Variational Splines and Paley-wiener Spaces on Combinatorial Graphs[J]. Constr Approx, 2009(29):1—21.
- [8] Craswell N, Hawking D. Overview of the TREC 2003 web track[C]// Text Retrieval Conference (TREC). 2003 Proceedings. [s. l.]:TREC, 2003.
- [9] Riget J, Vesterstroem J S. A Diversity-Guided Particle Swarm Optimizer—the ARPSO[EB/OL]. (2009)[2010-07-13]. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.2929>.
- [10] 焦巍, 刘光斌. 基于多样性反馈的粒子群优化算法[J]. 计算机工程, 2009, 35(22):202—204.
- [11] 徐文龙, 孙俊, 须文波. 基于干扰因子的 QPSO 算法改进[J]. 微电子学与计算机, 2008, 25(4):128—130.

作者简介:

高伟 男, (1981—), 博士, 讲师. 研究方向为统计学习理论及图论。

梁立 男, (1965—), 硕士, 教授. 研究方向为智能决策支持系统、信息检索。

作者简介:

范会联 男, (1971—), 硕士, 副教授, CCF 会员 (E20-0013523M). 研究方向为软件工程、智能信息处理。

陶维安 男, (1958—), 副教授. 研究方向为智能信息处理。

(上接第 14 页)