



基于图熵极值理论的领域概念聚类方法

安敬民¹, 李冠宇²

(1. 大连东软信息学院 计算机与软件学院, 辽宁 大连 116023; 2. 大连海事大学 信息科学技术学院, 辽宁 大连 116026)

摘 要: 为在领域本体学习过程中实现最优同领域概念聚类并解决概念重叠问题, 通过引入图熵极值理论, 提出一种新的领域概念聚类方法。依据最大信息熵原理, 将图中各概念节点视为一个整体以取代原选取质心的方法, 同时利用图熵最小化计算公式设计概念自动聚类机制。实验结果表明, 与 K-means 算法、基于密度和基于距离的领域概念聚类方法相比, 该方法可有效提高查准率、查全率以及综合评估指标 F 值。

关键词: 领域概念; 领域本体; 概念重叠; 图熵; 概念聚类

开放科学(资源服务)标志码(OSID):



中文引用格式: 安敬民, 李冠宇. 基于图熵极值理论的领域概念聚类方法[J]. 计算机工程, 2020, 46(6): 88-93.

英文引用格式: AN Jingmin, LI Guanyu. Domain concept clustering method based on graph entropy extreme value theory[J]. Computer Engineering, 2020, 46(6): 88-93.

Domain Concept Clustering Method Based on Graph Entropy Extreme Value Theory

AN Jingmin¹, LI Guanyu²

(1. School of Computer and Software, Dalian Neusoft University of Information, Dalian, Liaoning 116023, China;

2. Information Science and Technology College, Dalian Maritime University, Dalian, Liaoning 116026, China)

[Abstract] In domain ontology learning, in order to implement optimal clustering of concepts of the same domain without concept overlapping, this paper introduces the graph entropy extreme value theory and proposes a domain concept clustering method. According to the principle of maximum information entropy, the concept nodes of a graph are considered as a whole instead of selecting the centroid. Also, the graph entropy minimization formula is used to design an automatic concept clustering mechanism. Experimental results show that, compared with K-means algorithm, density-based and distance-based domain concept clustering methods, the proposed method significantly improves the precision, recall rate and comprehensive evaluation index, F value.

[Key words] domain concept; domain ontology; concept overlapping; graph entropy; concept clustering

DOI: 10.19678/j.issn.1000-3428.0054038

0 概述

在构建领域本体的过程中, 需要设计本体自学习机制以便在后期可以自动补充和扩展本体。而在本体自学习过程中, 同领域概念的聚类是关键步骤, 其决定所构建领域本体在实际应用过程中所提供领域概念的准确性。

同领域概念聚类的传统方法主要有基于划分、基于层次、基于形式概念分析以及基于语义距离的方法。基于划分的方法需要人为设定划分的个数和穷举所有可能的划分情况, 并给定递归算法将概念从一个划分移到另一划分来提高划分结果的准确性,

典型算法为 K-means^[1-2]。基于层次的方法是将复杂概念网中每一个概念节点视为一个独立的聚类, 利用领域一致度计算公式, 迭代合并同领域概念, 最终将复杂概念网分为多个领域的概念网, 其中基于分布密度计算^[3]的方法为主要代表。基于形式概念分析的方法主要采用概念格将对象分层, 模糊概念格^[4]和模糊 K-means^[5]是其中的代表方法。基于语义距离的方法主要通过计算领域概念在概念树中的距离进行聚类, 目前有基于遍历树的蚂蚁聚类算法^[6]和百科词条的本体概念聚类方法^[7]等。文献[1-2]采用优化后的 K-means 方法选取并优化聚类中心, 结合设定的聚类阈值实现同领域概念的聚

基金项目: 国家自然科学基金(61371090, 61602075); 辽宁省自然科学基金(20180550940)。

作者简介: 安敬民(1992—), 男, 讲师、硕士, 主研方向为智能信息处理; 李冠宇(通信作者), 教授、博士。

收稿日期: 2019-02-27 修回日期: 2019-07-02 E-mail: 870457569@qq.com。

类。文献[3]利用层次的耦合内聚比得到类数目的分布密度,通过密度聚类实现最终的概念聚类。文献[4-5]利用形式概念分析对概念的模糊处理能力,扩大了概念聚类范围,增强了聚类效果。文献[6]通过计算领域概念间的谷歌距离以及 Wikipedia 的距离和相似度实现聚类。文献[7]利用马氏距离得到概念向量间的语义距离,并通过多次迭代完成百科词条中的概念聚类。

上述方法是当前对于领域概念聚类处理的主要方法,但在聚类过程中均未考虑概念重叠问题,即一个概念(节点)可能同时属于多个领域,如“古董”一词,可以理解为“陈旧的事物”,同时也可以理解为“具有守旧思想的人”,因此,其同时属于两个领域的概念,而因为基于 K-means 划分和基于距离的方法所产生的概念集合总是不相交的,所以无法解决概念重叠问题。基于计算密度的方法和形式概念分析可以解决该类问题,但前者需要先选择初始种子概念作为聚类核心点,结合各概念周围的密度和设定的阈值,递归纳入新的概念并迭代循环执行(根据阈值会不同程度地不同聚类结果中出现相同概念词汇),而在此过程中有较多的主观因素涉及其中(如选取初始概念),也是影响聚类效果的主要原因;后者虽然可以通过隶属模糊和上下位近似的方法保留更多的概念信息,解决概念重叠问题,但同样需提供若干个选定的聚类中心,仍然受到人为因素的影响。

本文提出以图熵最大化实现初始节点随机选择代替质心法,以图熵最小化保证聚类结果的准确性。基于中文 WordNet^[8] 多角度计算概念间的语义相似度同时构建相似度矩阵,并将其转化为以概念为节点的无向图,利用图熵最小化计算公式^[9] 结合最大信息熵理论^[10],使图中概念节点能够在无需选取聚类质心的情况下实现最优同领域概念聚类,同时解决领域概念重叠问题。

1 问题描述与定义

现代汉语中有许多词语具有一词多义的特点,即一个词语同时具有多个领域概念。所以,在对该类词语进行同领域概念聚类时,其结果应同时出现在不同领域,如图1所示(以概念领域 D_i 和 D_j 以及其中的概念 $c_1 \sim c_8$ 和 c_k 为例),多个领域在概念上具有交集,而目前的概念聚类方法并不能很好地适用于该问题的处理(关于概念重叠问题的具体解释,此处不再赘述)。

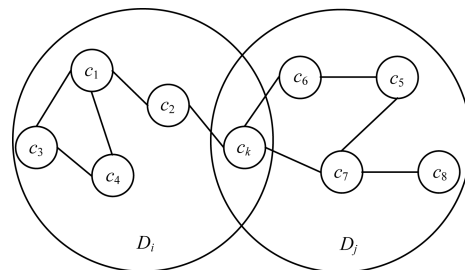


图1 领域概念聚类重叠示意图

Fig.1 Schematic diagram of domain concept clustering overlapping

从非结构化数据(如文本文档等)角度出发,对某个文本文档中的概念进行同领域聚类时,若 c_k 被归类在 D_i 中,则 c_k 不会出现在 D_j 中,造成 D_j 中的概念缺失而使查全率和查准率降低,本文针对此问题进行研究。

定义1 c_k 表示某个概念的词语,若 $\exists c_k$ 使得 $c_k \in D_i$ 且 $c_k \in D_j, i \neq j$, 则对 c_k 进行概念归类时会产生概念重叠现象。

2 基于 WordNet 的概念相似度综合计算

在对同领域概念进行聚类之前,要将概念间的相似度量化,根据概念与概念间的相似度判断聚类的方式,同时相似度的精度也直接影响聚类结果的准确性。本文通过结合中文 WordNet 多角度计算概念间的相似度,提高相似度计算的准确性。

2.1 概念间相似度计算方法

目前基于 WordNet 计算概念间相似度的方法有从概念在 WordNet 中的语义距离^[11]、深度及密度^[12]角度出发,或者考虑概念的语义重合度和概念包含的语义信息内容^[13]等方面相似度。本文结合文献[14]的设计思想,对各角度语义相似度计算方式进行综合考量,以提高概念相似度的精度。

基于 WordNet 中语义距离 $D(c_1, c_2)$ 的概念间相似度表示为 $C_s(c_1, c_2)$, 计算公式如式(1)所示:

$$C_s(c_1, c_2) = \frac{a}{a + D(c_1, c_2)} \quad (1)$$

其中, a 是 c_1, c_2 对中文 WordNet 中概念集合的平均语义距离。

以 $H_{\max}(c_1)$ 和 $H_{\max}(c_2)$ 分别表示 c_1, c_2 所在语义树的最大深度, $H(c_1)$ 和 $H(c_2)$ 分别表示 c_1, c_2 在树中的深度,则基于 WordNet 中语义树及概念深度的概念间相似度计算公式如式(2)所示:

$$H_s(c_1, c_2) = \frac{H(c_1) + H(c_2)}{H_{\max}(c_1) + H_{\max}(c_2)} \quad (2)$$

用 $n(c)$ 表示语义树中以 c 为根节点的直接子节点数, $n_{\max}(O)$ 表示与 c 所在同一语义树 O 中的所

有节点的直接子节点最大值,结合文献[12,14]计算概念相似度的方法,给出基于 WordNet 中语义树及概念密度的概念间相似度计算公式如式(3)所示:

$$P_s(c_1, c_2) = \frac{\sqrt{n(c_1)n(c_2)}}{n_{\max}(O)} \quad (3)$$

若将从 c 出发到语义树根节点所经过的节点个数记为 $S(c)$,则 c_1, c_2 基于 WordNet 语义树的语义重合度计算公式如式(4)所示:

$$S_s(c_1, c_2) = \frac{|S(c_1) \cap S(c_2)|}{|S(c_1) \cup S(c_2)|} \quad (4)$$

将式(4)结合文献[15-16]中的概念信息内容相似度计算方法,得到式(5):

$$I_s(c_1, c_2) = \frac{2I_{\max}(c)}{I(c_1) + I(c_2)} \quad (5)$$

其中, $I(c)$ 的计算公式见文献[13-14]。

2.2 概念间相似度综合计算

为提高概念间相似度计算的精确度,本文从多角度考虑对概念相似度有影响的因素,并将其以权重加和的形式作为综合计算结果。为使综合计算结果更具客观合理性,引入主成分分析法^[17]代替人为设定影响因子的方式。

构建矩阵 $\alpha = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})^T, i = 1, 2, 3, 4, 5$ 。其中 x_i 由形如 $\beta = (C_s, H_s, P_s, S_s, I_s)$ 的向量组构成。利用主成分分析法将 α 降为一维矩阵 $\gamma = [x'_1, x'_2, x'_3, x'_4, x'_5]$,并利用降维过程中得到的特征值计算主成分贡献率(q_1, q_2, q_3, q_4, q_5),最终得到概念间相似度综合计算公式为:

$$K_{\text{SIM}}(c_1, c_2) = q_1 x'_1 + q_2 x'_2 + q_3 x'_3 + q_4 x'_4 + q_5 x'_5 \quad (6)$$

3 同领域相似概念结构图

3.1 概念相似度矩阵

根据本文 2.2 节中两个概念间的同领域综合相似度 K_{SIM} ,按照对应概念 c_i 和 c_j 两个维度来构建概念相似度矩阵:

$$R_{\text{SIM}}(c_i, c_j) = \begin{bmatrix} 0 & K_{\text{SIM}}(c_1, c_2) & \cdots & K_{\text{SIM}}(c_1, c_n) \\ K_{\text{SIM}}(c_2, c_1) & 0 & & K_{\text{SIM}}(c_2, c_n) \\ K_{\text{SIM}}(c_3, c_1) & K_{\text{SIM}}(c_3, c_2) & & \vdots \\ \vdots & \vdots & & K_{\text{SIM}}(c_{n-1}, c_n) \\ K_{\text{SIM}}(c_n, c_1) & K_{\text{SIM}}(c_n, c_2) & \cdots & 0 \end{bmatrix}$$

其中, $1 \leq i \leq n, 1 \leq j \leq n$ 。

在矩阵 R_{SIM} 中,对角线部分是同一概念的相似度值,有 $K_{\text{SIM}}(c_i, c_i) = 1$,但由于在实际的领域概念聚类过程中强调不同概念的聚类,因此该值并无应用意义。为简化聚类操作,令 $K_{\text{SIM}}(c_i, c_i) = 0$ 。

3.2 同领域相似概念的图构建

设定阈值 λ ,当矩阵 R_{SIM} 中 $K_{\text{SIM}}(c_i, c_j)$ 大于 λ 时,将该值重新设定为 1;反之,若小于 λ ,则设定为 0。经过化简后的矩阵 R'_{SIM} 为一个只含有元素 0、1 的矩阵。

将简化得到的概念相似度矩阵 R'_{SIM} 转化为概念间的关系无向图 $G(V, E)$,其中, V 表示图中的概念节点(顶点), E 表示概念节点间的边。若在相似度矩阵 R'_{SIM} 中 $K_{\text{SIM}}(c_i, c_j) = 1$,则概念 c_i, c_j 对应的顶点 V_i, V_j 间存在无向边 E_{ij} ;反之,若 $K_{\text{SIM}}(c_i, c_j) = 0$,则说明概念间不存在无向边。最终形成同领域相似概念的图结构,如图 2 所示(以 $c_1 \sim c_7$ 和 c_n 为例)。

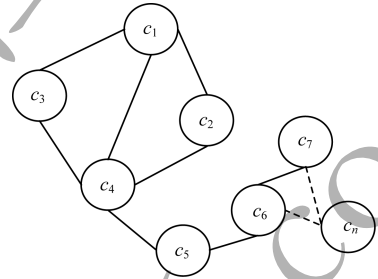


图 2 领域 D_i 中的相似概念结构图

Fig. 2 Similar concept structure graph in domain D_i

4 基于图熵的领域概念聚类优化

从信息论的最大信息熵理论角度出发,将图中的各个节点视为一个整体,而没有主客之分,即在本文的聚类过程中,区别以往传统的以聚类质心为主的聚类方法,而是将已聚类的节点作为整体,通过图熵最小化理论计算出下一步的最优聚类结果。利用图熵的聚类方法可以在每次聚类结果输出后保留原文本文档中已被聚类到某领域的概念,从而解决概念重叠问题。

设已构建的同领域相似概念结构图 $G(V, E)$ 的子图 $G'(V', E')$,对 G' 的内连接节点和外连接节点定义如下:

定义 2 在 $G'(V', E')$ 和 $G(V, E)$ 中, $\exists v_i \in V'$, $\exists v_j \in V$, 且 $\langle v_i, v_j \rangle \in E'$, 则 v_i, v_j 互为 G' 中的内连接节点; $\exists v_i \in V', \exists v_j \in V$, 且 $\langle v_i, v_j \rangle \notin E' \in E$, 则 v_j 为 v_i 在 G' 中的外连接节点。

在图 $G'(V', E')$ 和 $G(V, E)$ 中,点 v_i 的内连接率表示为 $P_{\text{in}}(v_i)$,外连接率表示为 $P_{\text{out}}(v_i)$,计算公式分别为:

$$P_{\text{in}}(v_i) = \frac{n}{N(v_i)} \quad (7)$$

$$P_{\text{out}}(v_i) = 1 - P_{\text{in}}(v_i) \quad (8)$$

其中, n 表示 v_i 的内连接节点数, $N(v_i)$ 表示 v_i 的内外连接节点数之和。

4.1 相似概念结构图的图熵处理方法

若设同领域相似概念结构图中某概念节点 c_i 的熵值^[18-19]为 $e(c_i)$, 结合信息熵^[20]公式和 c_i 的内外连接节点, 有:

$$e(c_i) = -P_{in}(c_i) \lg P_{in}(c_i) - P_{out}(c_i) \lg P_{out}(c_i) \quad (9)$$

文献[9]指出, 图熵可定义为所有节点的熵值之和, 即:

$$e(G) = \sum_{c_i \in V} e(c_i) \quad (10)$$

若每次聚类后得到的 $e(G)$ 最小, 则此时同领域相似概念聚类为最优。图3中虚线框内为经过若干步聚类后得到的最优聚类集合, 以加入 c_4 概念节点为例, 判断当前 $e(G)$ 的变化情况。结合上文公式可知: 加入 c_4 前, 计算得到 $e(c_1) = 0.92$, $e(c_2) = e(c_3) = 1$, $e(c_4) = 0.81$, $e(c_5) = e(c_6) = e(c_7) = e(c_n) = 0$, 则 $e(G) = 3.73$; 加入 c_4 后, $e(c_1) = e(c_2) = e(c_3) = 0$, $e(c_4) = 0.81$, $e(c_5) = 1$, $e(c_6) = e(c_7) = e(c_n) = 0$, 此时 $e(G) = 1.81$ 。由此可见: 加入 c_4 可使图熵减小, 聚类结果优化; 反之, 若使图熵 $e(G)$ 增大, 则拒绝加入 c_i 。

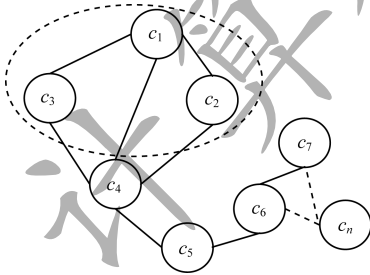


图3 聚类过程中图熵 $e(G)$ 变化示意图

Fig.3 Schematic diagram of change of entropy $e(G)$ in clustering process

4.2 基于图熵的领域概念聚类优化算法

从给定的图 G 中任意选取一个概念节点 c_i 作为起始点, 结合最大信息熵原理和图熵最小化计算公式, 最终形成一个满足同领域 D_i 最优概念聚类的子图 G' (概念集合), 具体算法如下:

输入 含有各领域概念的文本文档

输出 领域 D_i 的概念聚类集合表

1. 根据领域 D_i 给定的概念, 抽取文本文档中的同领域相似概念, 构建领域的相似概念结构图 $G(V, E)$

2. 令 $V' = \emptyset$

3. For 在 G 中任意选取一个节点 $c_i \notin V'$ 作为聚类起始点 do

4. { 遍历 c_i 的所有邻居节点, 并与 c_i 形成一个聚类子图 G'

5. For G' 中的 $V' \neq \emptyset$ do

6. { If 删除 G' 中的 c_j 后 $e(G')$ 变小

7. $G'. Delete(c_j)$

8. Else Continue }

9. For G' 外存在 V' 的邻居节点 do

10. { If 加入 G' 外的 c_k 后 $e(G')$ 变小

11. $G'. Add(c_k)$

12. Else Continue }

13. Return List(G') }

在给定具有各领域概念的文本中, 利用文献[2]提出的领域概念相关度和一致度计算公式抽取领域概念, 并结合概念相似度综合计算公式, 构建同领域相似概念的相似度矩阵, 并将其转化为对应的图关系, 通过图熵理论优化聚类结果。由于在此过程中聚类的对象是由文本文档构建的同领域相似概念结构图, 因此聚类后原文档中的概念并未删除, 使再次聚类某领域概念时仍可以抽取已被聚类过的概念, 从而解决领域概念重叠问题。

该算法首先任选图 G 中的某个节点 c_i 并遍历其邻居节点构建子图 G' , 此过程时间复杂度为 $O(n)$; 然后计算影响 G' 中图熵值增大的节点, 并做删除操作, 此过程时间复杂度为 $O(n \times m)$; 最后增加 G' 外使图熵减小的邻居节点, 此过程时间复杂度为 $O(n \times p)$ 。所以, 算法总的时间复杂度为 $O(n \times m + n \times p)$, 即 $O(n^2)$ 。

5 实验与结果分析

本文设计基于图熵极值化理论的领域概念聚类方法, 旨在对传统概念聚类方法在聚类查准率 (Precision) 上的优化以及对概念重叠问题的处理, 从而提高查全率 (Recall)。所以, 本文从领域概念的查准率和查全率以及综合评估指标 F 值 3 个方面将其与传统领域概念聚类方法进行对比。

本文实验环境是在 Windows10 下搭建的实验平台, 主要包括 Microsoft.NET framework 和 SQL Server 2012 database, 使用 C# 语言实现。选择由全国科学技术名词审定委员会审定公布的领域概念集合 (<http://www.cnctst.cn/>) 作为实验数据集, 从中选取 12 个不同的领域概念各 500 个, 将其混合后再以逐次增加数据的方式分为 4 组实验数据集, 分别为 800 个、1 200 个、1 800 个和 2 200 个领域概念。

由于本文贡献点在于处理中文领域概念聚类和聚类过程中的概念重叠问题, 因此与文献[1, 3, 7]方法在领域概念查准率、查全率以及综合评估指标 F-Measure 值 3 个方面进行对比, 并使用文献[2]中的计算公式, 如式(11)~式(13)所示:

$$\text{Precision} = \frac{N}{M} \quad (11)$$

$$\text{Recall} = \frac{N}{A} \quad (12)$$

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

其中, N 为聚类后属于领域 D_i 的概念个数, M 为聚类后领域 D_i 中的概念个数, A 表示所有文档的概念中属于 D_i 的概念个数。

通过实验获得的聚类结果以及 Precision 和 Recall 指标计算公式, 得到表 1 和表 2 所示的对比结果。

表 1 4 种方法的查准率对比

Table 1 Precision comparison of four methods

概念数	文献[1] 方法	文献[3] 方法	文献[7] 方法	本文方法
800	0.721	0.754	0.775	0.805
1 200	0.702	0.729	0.747	0.789
1 600	0.646	0.705	0.709	0.762
2 400	0.562	0.654	0.668	0.729

表 2 4 种方法的查全率对比

Table 2 Recall comparison of four methods

概念数	文献[1] 方法	文献[3] 方法	文献[7] 方法	本文方法
800	0.686	0.703	0.723	0.764
1 200	0.673	0.690	0.715	0.757
1 800	0.616	0.641	0.664	0.733
2 200	0.531	0.601	0.615	0.708

为减小实验比较误差, 将表 1 和表 2 中的查准率与查全率数据结合式(13)计算出 4 种方法在混合领域概念聚类上的 F 值, 比较结果如图 4 所示。

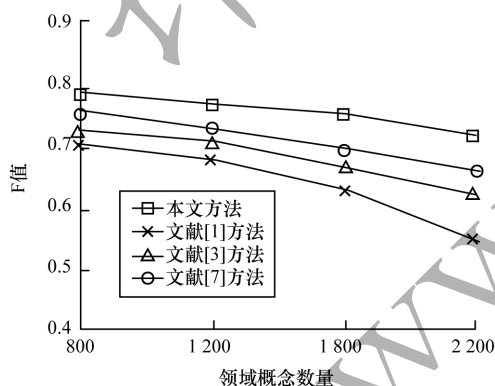


图 4 4 种方法的 F 值对比

Fig. 4 F value comparison of four methods

从图 4 可以看出, 4 种方法在第一次实验时, 相互间的 F 值差距不明显, 最低的文献[1]方法为 0.703, 而最高的则是本文方法达到 0.784。随着实验的进行, 本文方法相比其他方法优势逐渐明显, 在数据量为 2 200 的第 4 次实验中, 本文方法优势最为明显, 与基于传统 K-Means 划分方法的文献[1]方法相比 F 值提高近 30%, 同时基于密度计算的文

献[3]方法和基于距离计算的文献[7]方法相较于本文方法 F 值下降幅度也较为明显。

经分析每组实验数据可知, 随着概念数据量增加, 概念重叠情况也随之增多, 文献[1]方法对于该问题无法处理, 文献[3, 7]方法对其稍有作用, 但文中并未提及和考虑该问题, 没有提出有针对性的解决方案。本文方法在概念聚类过程中能够在保证原有聚类效率的前提下解决重叠问题, 提高了概念聚类性能。

6 结束语

本文针对领域概念聚类过程中的概念重叠现象, 提出基于图熵极值理论的领域概念聚类方法。利用图熵原理和在生成概念聚类的过程中不删除原文本领域概念的方法, 优化聚类结果, 提升查全率、查准率及 F 值。实验结果验证了该方法的有效性。但本文方法中阈值的选取仍为人工设定, 受到主观因素影响, 因此, 下一步将研究不同阈值的选取对聚类性能的影响, 并设计阈值自动生成机制。

参考文献

- [1] XU Dezhi, Junaid. An ontology learning based on documents clustering[J]. Computing Technology and Automation, 2010, 29(3): 49-52. (in Chinese)
徐德智, Junaid. Cluster-Merge 本体构造算法[J]. 计算技术与自动化, 2010, 29(3): 49-52.
- [2] MA Chuanbin. Research on key techniques of ontology learning based on Chinese text[D]. Xi'an: Xi'an University of Posts & Telecommunications, 2016. (in Chinese)
马传宾. 基于中文文本的本体学习关键技术研究[D]. 西安: 西安邮电大学, 2016.
- [3] HE Lin, HOU Hanqing. Research on semi-automatic construction of domain ontology based on statistical NLP technique[J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(2): 201-207. (in Chinese)
何琳, 侯汉清. 基于统计自然语言处理技术的领域本体半自动构建研究[J]. 情报学报, 2009, 28(2): 201-207.
- [4] QUAN T T, HUI S C, FONG A C M, et al. Automatic fuzzy ontology generation for semantic Web[J]. IEEE Transactions on Knowledge & Data Engineering, 2006, 2(3): 155-164.
- [5] KUMAR C A, SRINIVAS S. Concept lattice reduction using fuzzy K-means clustering[J]. Expert Systems with Applications, 2010, 37(3): 2696-2704.
- [6] WONG W, LIU W, BENAMOUN M. Tree-traversing ant algorithm for term clustering based on featureless similarities[J]. Data Mining & Knowledge Discovery, 2007, 15(3): 349-381.

- [7] YU Juan, CAO Xiao. Ontology concepts clustering based on encyclopedia entries[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(3): 636-640. (in Chinese)
于娟, 曹晓. 基于百科词条的本体概念聚类方法研究[J]. 电子科技大学学报, 2017, 46(3): 636-640.
- [8] WANG Shi, CAO Cungen, PEI Yajun, et al. A collocation-based method for semantic similarity measure for Chinese words[J]. Journal of Chinese Information Processing, 2013, 27(1): 7-14. (in Chinese)
王石, 曹存根, 裴亚军, 等. 一种基于搭配的中文词汇语义相似度计算方法[J]. 中文信息学报, 2013, 27(1): 7-14.
- [9] SHI Huanhuan, YIN Antao. An algorithm based on entropy clustering figure of overlapping community discovery[J]. Wireless Internet Technology, 2016(13): 98-101, 116. (in Chinese)
施欢欢, 印安涛. 基于图熵聚类的重叠社区发现算法[J]. 无线互联科技, 2016(13): 98-101, 116.
- [10] CHAI Lihe. Scientific and philosophical explanations on world vision from perspective of informationism[J]. Chinese Journal of Systems Science, 2014, 22(1): 21-25. (in Chinese)
柴立和. 信息主义视野下世界图景的科学及哲学诠释[J]. 系统科学学报, 2014, 22(1): 21-25.
- [11] LIU Feng, GUO Weiwei. A calculation model for research on concept similarity computing based on domain ontology[J]. Journal of Qufu Normal University (Natural Science), 2015, 41(4): 55-59. (in Chinese)
刘锋, 郭维威. 一种优化的基于领域本体语义距离的概念相似度计算模型研究[J]. 曲阜师范大学学报(自然科学版), 2015, 41(4): 55-59.
- [12] LIU Qianqian. Research on the calculation method of concept similarity in ontology mapping[D]. Wuhan: Wuhan University of Technology, 2014. (in Chinese)
刘茜茜. 本体映射中概念相似度计算方法的研究[D]. 武汉: 武汉理工大学, 2014.
- [13] HE Yan, ZHOU Zili. Concept IC model in WordNet based on entropy[J]. Computer Engineering, 2013, 39(10): 236-240. (in Chinese)
何艳, 周子力. 基于熵的 WordNet 概念 IC 模型[J]. 计算机工程, 2013, 39(10): 236-240.
- [14] WANG Tong, WANG Lei, WU Jiye, et al. Semantic similarity calculation method of comprehensive concept in wordnet[J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36(2): 98-101. (in Chinese)
王桐, 王磊, 吴吉义, 等. WordNet 中的综合概念语义相似度计算方法[J]. 北京邮电大学学报, 2013, 36(2): 98-101.
- [15] LIN Dekang. An information theoretic definition of similarity[C]//Proceedings of the 15th International Conference on Machine Learning. San Francisco, USA: Morgan/Kaufmann Publishers, 1998: 1-5.
- [16] NUNO S, TONY V, HAYES J. An intrinsic information content metric for semantic similarity in WordNet[C]//Proceedings of ECAI' 04. Valencia, Spain: [s. n.], 2004: 1-5.
- [17] FAN Xueli, FENG Haihong, YUAN Meng. PCA based on mutual information for feature selection[J]. Control and Decision, 2013, 28(6): 915-919. (in Chinese)
范雪莉, 冯海泓, 原猛. 基于互信息的主成分分析特征选择算法[J]. 控制与决策, 2013, 28(6): 915-919.
- [18] BAI L, ROSSI L, CUI L, et al. A novel entropy-based graph signature from the average mixing matrix[C]//Proceedings of International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1339-1344.
- [19] DAS K, DEHMER M. A conjecture regarding the extremal values of graph entropy based on degree powers[J]. Entropy, 2016, 18(5): 183-190.
- [20] ZHANG Xiao, MEI Changlin, CHEN Degang, et al. Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition, 2016, 56(1): 1-15.

编辑 金胡考