

面向语义信息查询的模糊本体模型

杨青, 陈薇, 闻彬

(华中师范大学计算机科学系, 武汉 430079)

摘要: 针对领域知识建模时的模糊性、不确定性与信息查询时的局限性, 提出一种基于模糊控制规则的模糊本体模型。利用基于模糊聚类的本体机器学习方法构建模糊控制规则库, 通过计算模糊相似矩阵得到模糊概念的语义关联, 对词汇相关概念进行语义分析与扩展获取模糊概念间的本质语义关系, 实现基于模糊概念属性值的信息查询与语义共用。实验结果表明, 该模型在语义查询上有更完善的推理机制, 能有效获取语义信息。

关键词: 模糊本体; 规则库; 语义查询; 模糊控制

Fuzzy Ontology Model for Semantic Information Query

YANG Qing, CHEN Wei, WEN Bin

(Department of Computer Science, Huazhong Normal University, Wuhan 430079)

【Abstract】 Aiming at fuzziness and indefinability for domain knowledge modeling and limitation for information query, this paper proposes a fuzzy ontology model based on fuzzy control rule. It constructs fuzzy rule base by using ontology machine learning method based on fuzzy clustering, gets semantic relation of fuzzy concept by computing fuzzy similar matrix. The intrinsic semantic relationship can be accessed for semantic analyzing and expanding of glossary relation concept, information query and semantic sharing are realized based on fuzzy concept attribute value. Experimental results show that this model has a better reasoning mechanism in semantic query, and it can get semantic information effectively.

【Key words】 fuzzy ontology; rule base; semantic query; fuzzy control

1 概述

本体作为共享概念模型的明确形式化规范说明^[1], 可获取、描述和表示相关领域的知识, 并提供对该领域知识的认识, 可理解或表达一个领域的概念。这些概念由实体、属性、相互关系和公理组成并明确定义。本体作为机器可读的形式存在, 被共同认可, 应用于很多不同领域, 例如知识抽取、信息查询等。但由于形式化的概念并不能够充分表达不确定的信息, 尤其是当前很多应用领域中存在大量不确定信息, 使领域概念之间缺少清晰的界限, 对于这些非确定性的知识, 纯粹利用概念描述的本体不能处理, 因此提出将模糊逻辑^[2]引入到本体中的研究。

文献[3]给出了一个模糊本体生成框架, 通过建立概念层, 将模糊逻辑整合到形式概念去处理不确定信息。文献[4]提出一个新的描述逻辑模糊扩展 FCDLS。这些常见的模糊本体模型主要利用普通模糊集来表达不确定性, 忽略了模糊概念间本质的语义关系, 这些本体模型无法处理作为属性值出现的本体概念。针对这些不足, 本文提出一种基于模糊控制规则库的模糊本体模型。

2 模糊领域本体的建立

模糊逻辑基于模糊集理论, 提供一种处理不确定信息的方式, 将隶属函数与集合中的元素对应起来, 使该元素与区间值[0,1]中的某个数相联系, 这个数就是相对于集合的隶属等级。 $X=\{x\}$ 表示对象空间, x 表示 X 的一般元素, 在 X 中的模糊集 F 可以表示为: $F=\{x, \mu_F(x) | x \in X\}$, 其中, $\mu_F(x)$ 表示 x 在模糊集 F 中的隶属等级。当 $\mu_F(x)=0$ 或 $\mu_F(x)=1$ 时, F 是非模糊集; 当 $\mu_F(x)$ 在0~1之间取值时, F 是一个模糊集,

此时 x 是否属于 F 不是一个明晰的概念。

模糊领域本体是传统的领域本体扩展, 在各领域本体概念属性中加入了隶属度集, 模糊领域本体可通过领域本体定义。

定义1 模糊领域本体是一个六元组:

$$O_F = \{C_x, C_i, P_F, O, R, A_C\}$$

其中, C_x 是对论域概念的描述, 是一个模糊的概念; C_i 是模糊概念 C_x 的集合; $R \subseteq C_i \times C_i$ 是概念关系集, 可以表示概念间相交、不相交、包含、补等基本关系, 也可表示概念之间模糊的关系; P_F 是概念属性集, 是一个四元组 $P_F(C_p, V_p, U, \mu_p)$, C_p 是本体概念集, V_p 是模糊属性值; $U = \{x_1, x_2, \dots, x_i\}$ 是论域, 假设模糊集合 A 用 $\{(x_1, \mu_p(x_1)), (x_2, \mu_p(x_2)), \dots, (x_i, \mu_p(x_i))\}$ 表示; $\mu_p: U \rightarrow [0,1]$ 是模糊集合的隶属度函数, $\mu_p(x_i)$ 表示模糊集合 A 中 x_i 的隶属程度; O 是 C_i 上的二元运算集合, 基本操作包括交运算、补运算、并运算等, 用于模糊本体概念的创建与合成; A_C 是公理的集合, 包括演绎规则和反应式规则, 用于对知识的推理或演绎, 用来处理知识库。

模糊领域本体通过领域专家的指导, 融合模糊逻辑理论构建, 模糊领域本体的生成如图1所示。

基金项目: 国家社会科学基金资助项目(07BYY033); 湖北省科技攻关计划基金资助项目(2007AA101C49)

作者简介: 杨青(1965-), 女, 副教授, 主研方向: 人工智能, 数据库, 语义网; 陈薇、闻彬, 硕士

收稿日期: 2009-10-17 **E-mail:** aillenwei@163.com

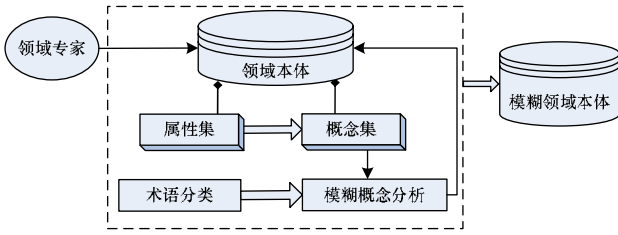


图1 模糊领域本体的生成

模糊领域本体用来处理特定领域的模糊信息，但对于基于属性值的模糊概念，不能很好地表达模糊概念间的语义关系和相关程度，特别是在不同本体概念之间，模糊语义的整合有很大的局限性，而且对同一个属性集，在不同的概念中可能有不同的取值，为了让属性值保持语义相关性，本文进一步构建基于模糊控制规则库的模糊本体模型。

3 模糊控制规则库的建立

模糊控制规则库的建立分为2个步骤：(1)在已经建立好的领域知识中提取重要概念，形成信息词汇集。(2)通过计算概念的语义模糊相似性，利用模糊概念聚类对信息词汇集中相关概念进行语义分析与扩展，制定概念之间的基本关系或模糊关系。

3.1 信息词汇集的建立

知识在模糊集理论中一个重要的考虑因素是将数据处理为语言^[5]，依据该分析方法，利用语言变量提供数据分析的一般特征建立信息词汇集。

定义 2 语言变量是在指定论域下描述自然语言的模糊集合，为了用逻辑推理描述代替自然语言描述，语言变量由一个五元组 $(x, T(x), U, G, M)$ 描述其特性，其中， x 代表变量名，如年纪、颜色等； $T(x)$ 表示 x 的词集合(term set)； U 代表涵盖变量范围的论域； G 代表建立语义变量 x 的语法规则； M 是连接每个词 x 的词义规则，以 $M(x)$ 表示论域 U 的模糊子集。

在信息词汇集中，各概念以语言变量的形式存储，用模糊语言值表示模糊概念内涵，其中包括概念的描述、解释以及实例，对于每个概念，至少存在一个描述。

3.2 模糊概念聚类

模糊聚类分析以模糊概念为基础，通过聚类技术实现，是本体机器学习的一种有效方法。本文通过计算信息词汇集中以语言变量值存在的概念间模糊相似性，得到相应模糊相似矩阵聚类模糊概念。

3.2.1 模糊相似性计算

概念间的模糊相似性 $sim(c_1, c_2)$ 可以根据概念的语义相似度 $sim_{semantic}(c_1, c_2)$ 和描述相似度 $sim_{description}(c_1, c_2)$ 按式(1)计算：

$$sim(c_1, c_2) = w_{sema} sim_{semantic}(c_1, c_2) + w_{desc} sim_{description}(c_1, c_2) \quad (1)$$

其中， $w_{sema} + w_{desc} = 1$ ； $sim_{semantic}(c_1, c_2)$ 表示语义相似度，由基于实例的方法 $sim_{instance}(c_1, c_2)$ 和启发式规则 $sim_{heuristicRule}(c_1, c_2)$ 的方法按式(2)计算：

$$sim_{semantic}(c_1, c_2) = w_{ins} sim_{instance}(c_1, c_2) + w_{heu} sim_{heuristicRule}(c_1, c_2) \quad (2)$$

其中， $w_{ins} + w_{heu} = 1$ ；基于启发式规则的方法制定依据一些领域公理和领域专家定义的规则；基于实例的方法是计算一定量的实例在2个概念中出现的联合分布概率，其计算公式见式(3)：

$$sim_{instance}(c_1, c_2) = \frac{p(c_1, c_2)}{p(c_1, c_2) + p(c_1, c_2) + p(c_1, c_2)} \quad (3)$$

其中， $sim_{description}(c_1, c_2)$ 是描述相似度，指描述概念的属性或关系间的相似程度，由关系相似度和属性相似度按式(4)计算得出；通过基于实例的方法计算关系 r_{i_1} 和 r_{2_j} 的相似度 $Rsim(r_{i_1}, r_{2_j})$ 由式(5)得出；属性相似度计算根据属性类型不同有很多方法，本文只给出对于语言文字属性的计算方法，该方法是基于属性实例的，其计算见式(6)：

$$sim_{description}(c_1, c_2) = w_{att} sim_{attribute}(c_1, c_2) + w_{rel} sim_{relation}(c_1, c_2) \quad (4)$$

其中， $w_{att} + w_{rel} = 1$ 。

$$sim_{relation}(c_1, c_2) = \frac{\sum_{i=1}^n w_{relation}^i R_{sim}(r_{i_1}, r_{2_j})}{\sum_{i=1}^n w_{relation}^i R_{sim}} \quad (5)$$

$$sim_{attribute}(c_1, c_2) = \frac{\sum_{i=1}^n w_{attribute}^i A_{sim}(a_{1_i}, a_{2_j})}{\sum_{i=1}^n w_{attribute}^i A_{sim}} \quad (6)$$

3.2.2 模糊概念聚类的实现

通过概念间的模糊相似性计算，建立相应的模糊相似矩阵 $\Phi = (AF_L(c_i, c_j))_{n \times n}$ ，利用模糊相似矩阵对模糊概念聚类，算法具体思想如下：

(1)将所有概念作为节点保存，然后利用模糊相似矩阵 $\Phi = (AF_L(c_i, c_j))_{n \times n}$ 得到各概念组的相似关联度，将各关联值按值非递增顺序依次存入数组 $A[\text{MAXSIZE}]$ 中，去掉重复值。

(2)根据数组 $A[\text{MAXSIZE}]$ 元素值的大小，将关联值最大的概念节点对连接起来，令其权值为该关联值，若连接2个节点时出现回路，则不连接这2个节点。当所有关联值概念对全部连接完毕后，删除数组 $A[\text{MAXSIZE}]$ 中的最大值。

(3)重复步骤(2)，直到数组 $A[\text{MAXSIZE}]$ 中无数据。

(4)设定近似置信阈值 $\delta \in [0, 1]$ ，把权值小于 δ 的节点连线去掉，得到关于 δ 的模糊概念聚类。

4 基于模糊控制规则的模糊本体模型

基于模糊控制规则库的模糊本体模型以语言变量为基础，继承并扩展模糊领域本体。

定义 3 基于模糊控制规则库的模糊本体是一个七元组 $O_F(C_x, C_i, R_m, O, L, I_F, A_F)$ ，其本体结构如图2所示。

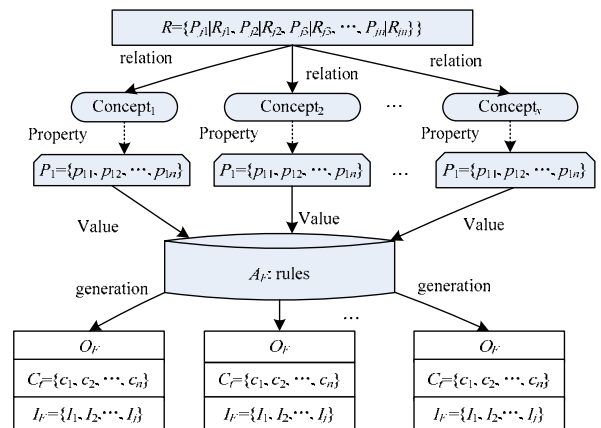


图2 基于模糊控制规则库的模糊本体结构

其中， C_x 是用模糊语言变量描述的抽象概念； C_i 是模糊概念 C_i 的集合； R_m 是概念关系集； O 是 C_i 上的二元运算集合， C_i, C_i, R_m, O 分别对应于定义1中的 C_i, C_i, R_m, O ； L 是模糊语言变量限定词的集合，与抽象的模糊语言变量结合组成

一定模糊强度的模糊概念； I_F 是模糊概念集 C_i 上实例的集合，实例也称为个体，是本体元素的实现； A_F 是模糊控制规则集，处理概念之间实质性的语义关系，限定各概念的语义扩展，同时包含概念的基本关系和模糊关系。

由于概念可以作为属性值出现，因此利用基于模糊控制规则库的本体模型来处理模糊信息，可以避免不同概念的属性值之间相同语义的排斥性，使模糊概念或属性概念之间有明确的语义关系，丰富本体的语义信息，有利于本体概念的共享与集成。

5 面向语义信息查询的模糊本体模型

对某个领域的具体描述通常基于一些模糊的语言。在教学本体中，“教师”有属性{教学态度，科研能力，授课水平，……}，其语言值{教学态度{很好，好，……，很差}，科研能力{很强，较强，……，弱}，授课水平{很高，高，……，低}等。作为“教师”属性出现的部分属性同时又作为“教学评价”的属性出现，即同一类属性出现在不同概念中，形成概念属性的共享。此外，在该领域本体中，本体概念“教师”能作为其他概念的属性。为了更有效地体现这些属性值与概念组之间的关系，本文将概念对应到模糊控制规则库中以语言变量的形式存储到词汇信息集中，利用模糊概念聚类分析得到概念间实质性的语义关系。

在输入检索词后，首先在模糊控制规则库中将模糊概念语义扩展，例如对于“优秀教师”的查询，即通过模糊控制规则库得到与概念“优秀”按 δ 水平聚类的所有模糊概念，并将其扩展到各属性，重新组合形成新的基于属性的模糊查询条件，再利用该模糊控制规则库的模糊本体模型创建基于属性值的查询。

设 $\{c_1, c_2, \dots, c_7\} \subseteq \{\text{概念词集}\}$ ，分别用来表示{优秀，好，…，低}，各概念之间的模糊相似性可以通过式(1)计算得到，下文以 (c_1, c_2) 为例进行说明。

由于概念无需具体的实例数据，因此置 $w_{ins} = 0$ ， $w_{att} = 0$ ，根据启发式规则法得到的语义相似性为

$$sim_{semantic}(c_1, c_2) = sim_{heuristicRule}(c_1, c_2) = 0.9$$

根据关系相似度计算描述相似性为

$$sim_{description}(c_1, c_2) = sim_{relation}(c_1, c_2) = 1$$

设 $w_{sema} = w_{desc} = 0.5$ ，概念的模糊相似性为

$$sim(c_1, c_2) = w_{sema} \cdot sim_{semantic}(c_1, c_2) + w_{desc} \cdot sim_{description}(c_1, c_2) = 0.95$$

依次计算 $\{c_1, c_2, \dots, c_7\}$ 中各概念对的模糊相似性，得到相应的模糊相似矩阵为

$$\Phi = \begin{pmatrix} 1.00 & 0.95 & 0.00 & 0.50 & 0.95 & 0.95 & 0.00 \\ 0.95 & 1.00 & 0.00 & 0.50 & 0.80 & 0.80 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.35 & 0.00 & 0.00 & 0.90 \\ 0.50 & 0.50 & 0.35 & 1.00 & 0.50 & 0.50 & 0.35 \\ 0.95 & 0.80 & 0.00 & 0.50 & 1.00 & 0.80 & 0.00 \\ 0.95 & 0.80 & 0.00 & 0.50 & 0.80 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.90 & 0.30 & 0.00 & 0.00 & 1.00 \end{pmatrix}$$

得到的概念连通示意图如图 3 所示。

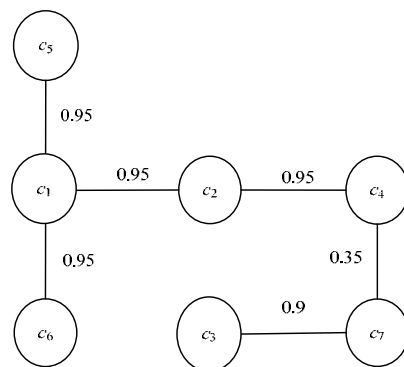


图 3 概念连通示意图

如果阈值 $\delta \in [0.9, 0.95)$ ，则聚类为 3 个类，分别为 $\{c_1, c_2, c_5, c_6\}$ ， $\{c_3, c_7\}$ ， $\{c_4\}$ 。阈值的选取根据查询的精度设定，阈值越高，类型间各概念的模糊相似性越大。

若选取 δ 为 0.9，利用模糊控制规则库对查询条件“优秀教师”语义扩展，得到的查询条件为“优秀教师 —属性→ 科研能力 —属性值→ 强”，“优秀教师 —属性→ 教学态度 —属性值→ 好”，“优秀教师 —属性→ 授课水平 —属性值→ 高”。同时，各概念在模糊控制规则库中还存在基本概念间的序关系，例如“优秀教师 —属性→ 科研能力 —属性值→ 强”，根据模糊控制规则库中的语义包含规则，推理出查询条件为“优秀教师 —属性→ 科研能力 —属性值→ 强”或“优秀教师 —属性→ 科研能力 —属性值→ 较强”或“优秀教师 —属性→ 科研能力 —属性值→ 很强”，因此，基于模糊控制规则库的模糊本体模型既能保持查询的精度，又能防止信息漏选。

6 结束语

为了提高语义信息处理的准确性和有效性，自然语言中可能出现以概念属性作为查询条件，本文在本体模型中构建一种模糊控制规则库，提出一个面向语义信息查询的基于模糊控制规则库的模糊本体生成模型。该模型能够完成智能化的信息检索。下一步的研究方向包括模糊本体的自动生成，模糊本体生成的一致性检测等。

参考文献

- [1] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.
- [2] Zadeh A. Similarity Relations and Fuzzy Orderings[J]. Information Science, 1970, 3(2): 177-200.
- [3] Quan Thanh Tho, Hui Siu Cheung, Cao Tru Hoang. FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web[C]//Proc. of KDO'04. Pisa, Italy: [s. n.], 2004.
- [4] Kang Dazhou. Description Logics for Fuzzy Ontologies on Semantic Web[J]. Journal of Southeast University, 2006, 22(3): 343-347.
- [5] Ling Feng, Tharam S D. Using Fuzzy Linguistic Representations to Provide Explanatory Semantics for Data Warehouses[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(1): 86-102.

编辑 陆燕菲