基于本体的构件聚类研究

张 锋，张英俊，潘理虎，谢斌红，陈立潮
（太原科技大学 计算机科学与技术学院，山西 太原 030024）

摘 要：为了消除自然语言对构件文本信息描述的二义性以及增强术语间的语义关系，文中采用领域本体的思想，给出了一个基于人工智能领域本体的软件构件聚类模型和基于该模型的聚类算法。该模型通过分析领域的共同概念，形成领域本体知识库，提供领域内一致认可的术语，用于匹配对构件文本描述所使用的自然语言。给出的算法通过与基于传统空间向量的K-Means算法分析比较，验证了该算法是有效的，实现了对软件构件更合理的聚类，提高了构件检索的效率和准确性。

关键词：聚类；本体；空间向量模型；本体语言
中图分类号：TP31 **文献标识码**：A **文章编号**：1673-629X(2012)11-0045-03

Research of Component Clustering Based on Ontology

ZHANG Feng, ZHANG Ying-jun, PAN Li-hu, XIE Bin-hong, CHEN Li-chao
(School of Computer Science and Techn., Taiyuan University of Science and Techn., Taiyuan 030024, China)

Abstract: For eliminating the ambiguity of component description originated from using natural language and enhancing the semantic relationship among terms, a software component clustering model based on artificial intelligent domain ontology and clustering algorithm based on this model was presented with domain ontology. The model has abstracted the domain knowledge formation ontology library, which was used to match what users input in natural language and outputs coherent retrieval terms in the domain. The experiments prove the method is effective, which implements a more rational classification of the components, and improves the efficiency and accuracy of component retrieval.

Key words: clustering; ontology; VSM; OWL

0 引 言

软件复用技术已成为当前软件工程界研究的热点，它能够缩短软件的开发周期以及提高软件产品的质量、避免程序员的重复劳动。随着IT技术的发展，软件构件技术已成为一门独立的学科[1]。构件库系统负责对构件的有效管理，包括构件的分类和构件的检索机制[2]，合理的构件分类可以提高构件的检索效率，因此，文中对构件的分类进行研究。

当前，构件的分类有多种表示方法，H. Mili从检索效果和复杂度的角度将其分为了基于词法描述子的、基于文本的和基于规约的编码与检索方法。W. Frakes从构件表示出发将其分为了人工智能方法、超文本方法和信息科学方法三类[3]。其中，W. Frakes信息科学方法中的刻面分类是复用项目中应用最为成功、较为广泛的一类构件表示方法。针对刻面分类表示法的术语空间依赖于专家经验，具有较强人为主观因素的不足[4]，文中采用结合领域本体的方法对构件进行聚类，通过增强术语间的语义关系，使得聚类结果更为合理。

本体已有广泛的应用，因此诞生了许多本体描述语言，如RDF[5]和RDF-S[6]，KIF，SHOE[7]，XOL[8]，OWL等，这里对构件的描述采用OWL语言。

1 本体的构件聚类过程

基于本体的构件聚类系统，主要包括三部分：本体构建，构件特征表示，构件聚类。具体的系统模型结构如图1所示。

在模型中，首先对要聚类的目标收集数据，得到一个构件库，由于每个构件是用文本描述的，因此对文本集需要进行预处理，即利用自动分词技术把文档集进行切分，进而可以将每一个文本表示为特征向量，然后根据领域本体对特征向量进行降维，从而为下一步的聚类做好准备，在整个聚类过程中，领域本体将起到至关重要的作用。

收稿日期：2012-03-05；**修回日期**：2012-06-11
基金项目：山西省自然科学基金(2009011022-1)；太原科技大学研究生创新项目
作者简介：张 锋(1985-)，男，硕士研究生，研究方向为人工智能；张英俊，高级工程师，硕士，研究方向为软件体系结构、Web Services。

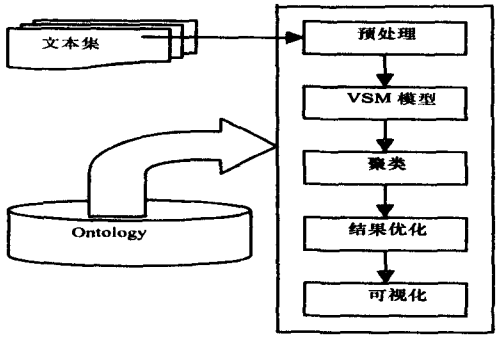

图1 模型结构

2 本 体

本体是对共享概念模型的明确的形式化规范说明[9]。在计算机领域,它的作用是获取通用知识,提供本领域内一致认可的理解。它形式化定义领域内的概念,同时也强调这些概念之间的联系[10]。

2.1 本体构建

本体领域模型主要体现的是领域内概念与概念之间关系的表示,机器能够自动推理。本体提供明确的术语定义,使得概念间的关系极易被机器理解进而得到处理。在文中用到的本体是根据上海构件库人工智能领域利用 Protégé 构建的,首先要列举出本领域内的所有概念,并对这些概念进行定义。人工智能领域构件的基本描述信息主要由功能类型、开发语言、运行环境、操作系统、开发状态、面向用户、许可证、国际化支持组成,并且他们都拥有各自的术语空间。比如构件功能类型有文本处理、科研和教育、程序开发、游戏和娱乐等术语;非功能类型主要包括开发语言、运行环境、操作系统等等。其次还要描述概念与概念之间的关系,比如开发语言包括 java、C、C++等,它们在本体中对应的就是 kind-of 关系。本体的构建如图2所示。

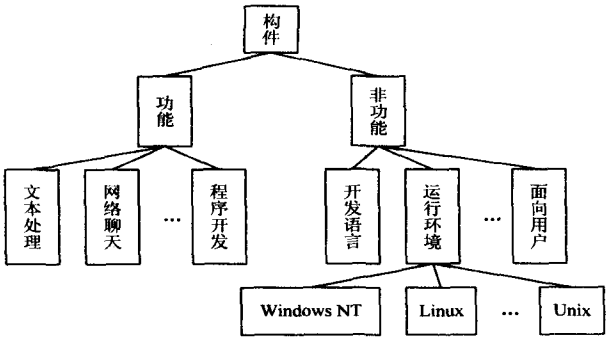

图2 本体层次结构

其中一个类的 OWL 描述如下:

```
<owl Class rdf:ID="Linux">
<rdfs:subClassOf rdf:resource="#运行环境"/>
</owl:Class>
```

2.2 本体匹配

将一个词汇的集合匹配到领域本体,就得到了公认的概念集,一个构件的文档集合可表示为 $T=\{t_1,t_2,\cdots,t_n\}$,将集合 T 与领域本体匹配后可以得到一个概念集 $C=\{c_1,c_2,\cdots,c_m\}$,领域本体采用 OWL 来描述,因此制定了如下的匹配规则[11]:

1) 当集合 T 中 $t_i(i=1,2,\cdots,n)$ 与本体中的概念可以匹配时,将匹配到的概念 $c_j(j=1,2,\cdots,m)$ 加入到集合 C 中。

2) 当集合 T 中 $t_i(i=1,2,\cdots,n)$ 不仅与本体中概念的属性可以匹配,还和别的概念匹配或该概念的实例匹配时,则遵循"概念的优先级最高,实例优先级次之,属性优先级最低"的原则,则将该概念或实例所对应的概念作为匹配的概念输出,否则,将该属性对应的概念加入到集合 C 中。

3) 当集合 T 中 $t_i(i=1,2,\cdots,n)$ 只与本体中的某个实例匹配时,则将其所属的概念加入到集合 C 中。

4) 当集合 T 中的元素不与任何概念匹配时,则将其丢弃。

由于现在的词都是经过与领域本体匹配后得到的公认概念,因此本文定义了两个概念之间相似度的计算公式:两概念 C_1 和 C_2,它们各自在本体中的属性为 $C_1=\{t_1,t_2,\cdots,t_m\}$;$C_2=\{t_1',t_2',\cdots,t_m'\}$,如果 $\{t_1,t_2,\cdots,t_m\}\cap\{t_1',t_2',\cdots,t_m'\}=\{p_1,p_2,\cdots,p_k\}$,同时定义集合 $p=\{p_1,p_2,\cdots,p_k\}$ 中每个概念的权重分别为 $\{w_1,w_2,\cdots,w_k\}$,则可以得到两概念之间的相似度为:

$$\mathrm{Sim}(c_1,c_2)=(w_1+w_2+\cdots+w_k)\div k \qquad (1)$$

通过得到两概念的共同属性,将这些共同属性的权重求和后取平均值,就得到了两个概念间的相似度。

2.3 基于本体的 VSM 表示模型描述

为了得到每个文本的一个特征向量,首先对文本集进行预处理,然后将得到的特征向量映射到领域本体上,从而得到了一个新的特征向量。

首先利用软件 IKAnalyzer 对文本集进行切分,每个文本即可表示为 $d_i(t_1,t_2,\cdots,t_n)$(其中 $i=1,2,\cdots,m$),然后将 $t_j(j=1,2,\cdots,n)$ 与领域本体匹配可以得到一个新的向量 $\mathrm{Ont\_d}_i=(c_1,c_2,\cdots,c_n)$。在传统的向量空间模型中,文本可以表示为一个如下的向量:

$$V(d)=(t_1,w_1(d);t_2,w_2(d);\cdots;t_n,w_n(d))$$

式中 t— 词条项,$w(d)$—t 在 d 中的权重。TF-IDF 是一种常用的词条权重计算方法,这里采用该方法来计算权重。一个文本可表示为 n 维空间的一个向量,称 $V(d)=(w_1,w_2,\cdots,w_n)$ 为向量空间模型。其中每个词条的权重计算如下:

$$\mathrm{tfidf}(d,t)=tf(d,t)\times\log\left(\frac{|D|}{df(t)}\right) \qquad (2)$$

式中 D— 文本集,d— 任意文本,t— 文本中的词,$tf(d,t)$— 词 t 在文本 d 中出现的频率,$|D|$— 文本集的

数目,$df(t)$— 词 t 在文本集中出现的次数,那么 tfidf(d,t) 就为词 t 在文本 d 中的权重。

2.4 本体语义扩展与推理

本体语义扩展主要是术语描述的扩展,利用领域模型本体中概念间的关系,将一个术语扩充为一个术语集合,为构件聚类提供语义推荐。

本体知识库为构件聚类提供了语义推理所必须的规则和条件,通过 OWL 对本体的形式化描述,计算机能够理解本体所描述的知识,进而可以从明确的知识中推导出隐含的知识。在知识库推理系统中,公理是通过子类、子属性、属性定义域、属性值域等来描述的。

2.5 基于本体的构件聚类算法

传统的 K-Means 算法需要预先设定聚类参数,即确定 K 值,K 的取值不同对聚类结果影响较大,文中的算法思想不依赖 K 的取值,具体做法:采用取第一个构件为聚簇点,然后依次计算各个构件与它的相似度。需要设定一个阈值 simArg,当 sim(d_i,c_j) ≥ simArg 时,就把它放在一个簇内,反之令这个构件为第二个簇点,把剩下的构件与簇点比较,依次类推。

算法的具体描述如下:

1) 初始化,利用分词技术把文本集中的构件描述信息分别进行切分并存储在构件数组中;

2) 加载规则,将规则库中的全部规则加载到内存中,规则库中的规则使用 xml 进行存储,规则采用四元组表示,即 $R(rf,r,rsf,rs)$ 其中 R 表示规则,rf 表示推理假设刻面,r 表示推理假设,rsf 表示推理结论刻面,rs 表示推理结论;

3) 本体语义扩展,根据加载的规则将术语进行语义扩展,使一个术语扩展为一个术语集合;

4) 本体推理,根据 Protégé 类的三大公理 SubClassOf、EquivalentClasses、DisjointClasses 计算机智能进行推理;

5) 取第一个构件为簇点 c_j,分别计算其它构件与它的相似度 sim(d_i,c_j),在这里我们设定一个阈值 simArg,当构件间的相似度 sim(d_i,c_j) ≥ simArg,则把它们聚在一个簇内,反之,令当前的构件为第二个簇点,分别计算其它构件与簇点的相似度,依次类推;

6) 聚类,最后把一些小簇(文中认为簇内构件数 ≤ 10) 合并在一起,对这个类进行单独处理。

由此可知,该算法最大的优点在于不需要预先设定聚类个数;但是聚类后可能会产生一些小类,例如一些类中可能只含有一个或几个构件,因此还需要对这些小类进行优化合并,才能得到更合理的聚类结果。

3 实验结果与分析

通过将基于本体的构件聚类算法与基于空间向量模型的 K-Means 算法进行分析与比较,实验结果证实了基于本体的构件聚类算法是有效的。通过聚类结果可以明显地看出:两个构件的概念描述并不相同却被聚在了一个簇内,这充分体现了本体在语义方面的作用。实验数据来自于上海构件库[12]领域相关的人工智能构件,采用聚类精度、聚类召回率以及 F-Score 系数三项来进行比较:

(1)聚类精度(P):

$$P = \text{Precision}(i,j) = \frac{N_{ij}}{N_j} \tag{3}$$

(2) 聚类召回率(R):

$$R = \text{Recall}(i,j) = \frac{N_{ij}}{N_i} \tag{4}$$

(3)F - Score 方法:

$$\text{F - Score} = \frac{2 * P * R}{P + R} \tag{5}$$

式中,N_{ij}— 聚类 j 中分类 i 的构件数量;N_j— 聚类 j 中的所有构件数量;N_i— 分类 i 中所有构件的数量。

在以上三个指标下,基于空间向量模型的 K-Means 算法的聚类精度约为 55%,聚类召回率约为 60%,F-Score 约为 58%,而基于本体的构件聚类算法在这三个指标下的数据分别约为 79%、70%、75%。两种构件聚类算法的对比情况如图 3 所示。从图 3 中可以看出,基于本体的构件聚类算法优于基于空间向量模型的 K-Means 算法,其聚类精度高出 24%,聚类召回率高出 10%,F-Score 系数高出 17%。实验结果可以看出,通过引入领域本体,能够提高构件聚类的质量;进而改善了构件检索的效率和准确性。

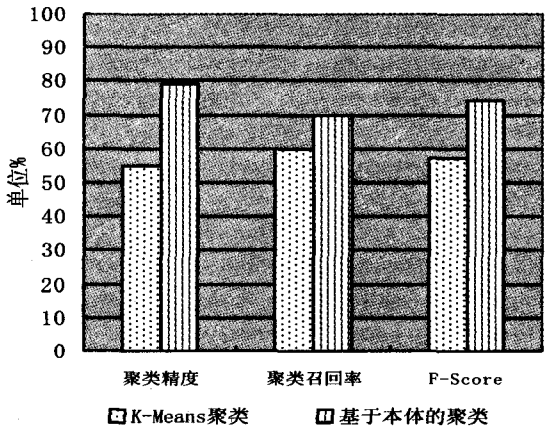

图 3　聚类效果对比

4 结束语

文中将领域本体与文本的聚类方法结合,对构件文本集进行聚类,通过对文本集进行预处理,得到每个文本的一个特征向量,然后再将它与领域本体匹配得

(下转第 52 页)

机器人结束行走，根据规划，此时的支撑腿 x 方向上恢复到初始位置，使得质心在 z 方向上升高，在 x 和 y 方向质心没有位移。在（2.5 ~ 3）秒时为结束行走的最后阶段，机器人各关节均恢复到了初始状态，机器人的质心在 y 方向恢复都零点，z 降低 ΔH。由图(d)，(e)利用三次多项式插值的方法使机器人各关节平滑，克服了关节数据突变的问题，保证了机器人运动过程的稳定[12]，并使得机器人在步行过程中其质心过渡平稳，实现一种静态的斜坡行走。

5 结束语

文章以机器人 Nao 为研究模型，将机器人行走过程划分为若干行走状态，刻画了一种机器人的步行流程图，用几何法对机器人斜坡行走的各关键状态进行了规划，克服了机器人行走中由于身体模型的改变给机器人稳定带来的影响，使机器人在行走过程中使得每个状态都能够保持机器人的静态稳定。最后利用三次多项式插值的方法对机器人关节的具体运动进行规划，得到平滑的关节运动轨迹，再对这些运动进行积分，获得平稳的质心运动轨迹，从而可以断定，此规划方法能够实现机器人在斜坡上的平稳行走。

参考文献：

[1] 许艳惠. 一种双足机器人的步态规划研究[J]. 核电子学与探测技术，2010，30(4)：542-545.

[2] 姜 山，程君实，陈佳品，等. 基于遗传算法的两足步行机器人步态优化[J]. 上海交通大学学报，1999，33(10)：1280-1283.

[3] 付成龙，陈 恳. 五杆四驱动平面双足机器人动态步态规划与非线性控制[J]. 机器人，2006，28(2)：206-212.

[4] Huang Q，Yokoi K，Kajita S，et al. Planning Walking Patterns for a Biped Robot[J]. IEEE Transactions on Robotics and Automation，2001，17(3)：280-289.

[5] Napoleon S N，Sampei M. Balance Control Analysis of Humanoid Robot Based on ZMP Feedback Control[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Lausanne，Switzerland：[s. n.]，2002：2437-2442.

[6] 李龙澍，王唯翔，王 凡. 基于三维线性倒立摆的双足机器人步态规划[J]. 计算机技术与发展，2011，21(6)：66-69.

[7] Kajita S，Kanehiro F，Kaneko K，et al. Biped Waling Pattern Generation by Using Preview Control of Zero-Moment Point [C]//Proceedings of the IEEE International Conference on Robotics and Automation. Taipei，Taiwan：[s. n.]，2003：14-19.

[8] 黄春林，张 祺，杨宜民. 三次样条差值方法在 Nao 机器人步态规划中的应用[J]. 机电工程技术，2011，40(2)：62-64.

[9] 肖 乐，常晋义. 仿人机器人下楼梯的自适应模糊控制方法[J]. 计算机工程，2009，35(13)：193-195.

[10] 王 凡，王 侠. RobotCup 仿真平台中 Nao 模型正运动学研究[J]. 合肥师范学院学报，2011，29(3)：49-51.

[11] 张 博，杜志江，孙立宁，等. 双足步行机器人步态规划方法研究[J]. 机械与电子，2008(4)：52-55.

[12] 谭 民，徐 德，侯增广，等. 先进机器人控制[M]. 北京：高等教育出版社，2007.

（上接第47页）

到每个文本的概念向量模型，这样能够有效地提高聚类速度以及聚类精度。在研究中还存在一些不足，比如本体的构建过程是手工参与的，在概念的匹配方面还需要找到更加有效的量化方法，这些问题将是下一步研究的内容。

参考文献：

[1] 杨芙清，梅 宏，李克勤. 软件复用与软件构件技术[J]. 电子学报，1999，27(2)：68-74.

[2] 王渊峰，薛云皎，张 涌，等. 刻面分类构件的匹配模型[J]. 软件学报，2003，14(3)：401-408.

[3] Frakes W B，Pole T P. An Empirical Study of Representation Methods for Reusable Software Components[J]. IEEE Transactions on Software Engineering，1994，20(8)：617-630.

[4] 常继传，李克勤，郭立峰. 青鸟系统中可复用软件构件的表示与查询[J]. 电子学报，2008，28(8)：20-23.

[5] Beckett D，Mcbride B. RDF/XML Syntax Specification[EB/OL]. 2004-02-10. http://www.w3.org/tr/rdf-syntax-grammar/.

[6] Brickley D，Guha R V. RDF Vocabulary Description Language 1.0：RDF Schema[EB/OL]. 2004-02-10. http://www.w3.org/tr/rdf-schema/.

[7] Heflin J，Hendler J. Searching the web with SHOE[C]//Proc of AAAI Workshop on AI for Web Search. [s. l.]：[s. n.]，2000：35-40.

[8] Karp P D，Chaudhri V K，Thomerel J. XOL：An XML-based Ontology Exchange Language[EB/OL]. 1999. http://www.oasis-open.org/cover/xol-03.html.

[9] Studer R. Knowledge Engineering，Principles and Methods [J]. Data and Knowledge Engineering，1998，25(12)：161-197.

[10] 吴 强. 语义 Web 中以描述逻辑为本体语言的推理[J]. 计算机工程与应用，2003，39(33)：30-32.

[11] 谢红薇，颜小林，余雪丽. 基于本体的 Web 页面聚类研究[J]. 计算机科学，2008，35(9)：153-155.

[12] 上海构件库[DB/OL]. 2011-09-26. http://www.sstc.org.cn/.