# 基于领域知识重用的虚拟领域本体构造*

陈 刚+, 陆汝钤, 金 芝

(中国科学院 数学与系统科学研究院,北京 100080)

# Constructing Virtual Domain Ontologies Based on Domain Knowledge Reuse

CHEN Gang+, LU Ru-Qian, JIN Zhi

(Academy of Mathematics and System Sciences, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: 86-10-62554389, E-mail: cg@amss.ac.cn

http://www.amss.ac.cn

Received 2001-10-12; Accepted 2002-04-10

**Chen G, Lu RQ, Jin Z. Constructing virtual domain ontologies based on domain knowledge reuse. *Journal of Software*, 2003,14(3):350~355.**

**Abstract**: A methodology is presented to construct virtual domain ontologies reusing other domain knowledge already available. This methodology uses the semantic relevance of existing domain models and domain ontologies, and proposes the possibility of building ontologies following the view of semantic match. Firstly, a structural definition of domain ontology is given. Secondly, the semantic relevance between domain model and ontologies is discussed and a definition of relevance degree is given. Based on the evolution of population, the domain ontologies constructing technologies in the terminology from genetics such as selection, clone, mutation, crossover, synthesis and transgenic are discussed. Finally, a VDO (virtual domain ontologies) constructing system is introduced, and a case study in the field of some hotels and travel agencies is demonstrated.

**Key words**: domain knowledge base; domain knowledge reuse; domain ontology; semantic relevant degree; ontology constructing

**摘 要**: 提出了一种重用现有领域知识库知识构造新领域本体的方法.该方法充分利用了领域知识模型以及领域本体相互之间存在的语义相关性,从语义匹配的角度探讨了构造新领域本体的可能性.首先给出了领域本体的一种结构化定义,然后讨论了领域模型之间、领域本体之间存在的语义相关性,并给出了领域本体语义相关度的概念.以此为基础,重点讨论了基于生物种群进化方法构造新领域本体的选择、克隆、变异、杂交、合成和转基因方法.最后详细介绍了一个虚拟领域本体构造系统,并给出了具体分析实例.

**关键词**: 领域知识库;领域知识重用;领域本体;语义相关度;本体构造

* Supported by the National Natural Science Foundation of China under Grant Nos.69983010, 60233010 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2001AA113130 (国家高技术研究发展计划); the National Grand Fundamental Research Pre-973 Program of China under Grant No.2001CCA03000 (国家重点基础研究发展规划前期研究专项(973)); the National Grand Fundamental Research 973 Program of China under Grant No.2002CB312004 (国家重点基础研究发展规划(973)); the CAS Project of Brain and Mind Science and the Knowledge Innovation Program of the Chinese Academy of Sciences of China (中国科学院知识创新工程)

**第一作者简介**: 陈刚(1968－),男,湖南永州人,博士,讲师,主要研究领域为基于知识的软件工程.

中图法分类号: TP311 文献标识码: A

有关本体论方法的研究和应用在知识工程、自然语言理解和知识表示等领域日益受到重视,特别是由于它在智能信息集成、Internet 信息获取和大规模知识库工程方面取得的成功,更使其成为人工智能界引人瞩目的热点之一.我们曾提出了面向本体的领域需求分析 OORA 方法[1~3],设计实现了领域描述语言 DODL 和领域分析语言 ONONET,以及基于本体的多领域知识库管理系统 DOKM[4].领域专家在我们的系统环境中,使用不含任何软件专业术语的描述语言 DODL 书写领域描述文档,构造生成新的领域本体.构造领域知识模型本身是一个相当繁琐、费时的工作,并且为了保持领域知识描述的一致性,领域知识库的维护工作也非常繁重.因此,如何利用知识库系统中已有的领域知识构造新的领域模型和领域本体,以实现领域知识重用成为一个亟待解决的问题.我们在 DOKM 系统中设计实现了虚拟领域本体 VDO(virtual domain ontology)子系统,其基本思想是,只在领域知识库中保存最基本的领域本体,当用户需要新的领域本体时,由系统根据用户提出的具体需求,对已有的领域本体进行组合或删减,动态地构造出新的领域本体.因为这些领域本体并不是由现实世界中提取出来,而是建立在其他领域本体基础之上的,所以称其为虚拟领域本体.本文第 1 节给出了领域本体的定义以及本体语义相关度的概念.第 2 节讨论了虚拟领域本体构造技术.第 3 节介绍了我们实现的原型系统,并给出了一个分析实例.第 4 节描述了本文的相关工作.

## 1 本体与本体语义相关性

### 1.1 领域本体的概念

领域本体是用于描述指定领域知识的一种专门本体.它给出了领域实体概念及相互关系、领域活动以及该领域所具有的特性和规律的一种形式化描述.我们认为,领域本体由属性、对象、关系和子领域本体组成[1],这意味着:(1) 领域实体间的关系和实体对象都是独立的知识单元;(2) 领域本体可嵌套,在领域本体中可以再嵌入其他领域本体(称为子领域本体).下面给出领域本体的形式化定义(定义中提及的对象和对象类的概念与面向对象方法中的同名概念完全相同).

**定义 1(无环超图)**. 一个无环超图表示为 $G=(D,P,U,V,H)$,集合 $D,P,U,V,H$ 互不相交.其中 $D$ 为描述符集;$P$ 为处理符集;$U$ 和 $V$ 为图 $G$ 顶点的有限集,分别称为简单顶点和复杂顶点,而复杂顶点本身又是一个超图;$H$ 是一个超平面有限集,该超平面集被惟一命名并由 $U\cup V$ 子集确定;$D\cup P\cup U\cup V\cup H$ 集中的元素都包含在超图 $G$ 中.

**定义 2(嵌套超图)**. 具有非空有限顶点集 $V$ 的无环超图称为嵌套超图.用 $d$ 表示嵌套超图$(D,P,U,V,H)$的嵌套深度,则 $d=1$,当 $V$ 中任何顶点描述的超图均没有嵌套时;$d=n+1$,当 $V$ 中顶点描述的超图最大嵌套深度为 $n$ 时.没有嵌套的超图,其嵌套深度为 0.对所有 $n\geq 0$,具有嵌套深度 $n$ 的超图被称为是有限嵌套的.

$U\cup V$ 中的两个顶点 $x$ 和 $y$ 被称为直接连接的,如果 $H$ 中的超平面 $h$ 满足 $x$ 和 $y$ 都属于 $h$.$U\cup V$ 中的每个元素都直接连接到它所属的超图.顶点 $x$ 和 $y$ 称为连接的,如果它们是直接连接的,或存在另一个顶点 $z$ 使 $x$ 与 $z$ 直接连接并且 $z$ 与 $y$ 直接连接.一个超图是连接的,如果图中所有顶点都是两两连接的.

**定义 3(本体)**. 本体是一个连接的、有限嵌套的超图,其中集合 $U\cup V$ 的势不小于 2.超图中的描述符集 $D$ 称为本体的属性,处理符集 $P$ 称为本体的方法,简单顶点 $U$ 称为本体中含有的对象,复杂顶点 $V$ 还是一个本体,超图 $H$ 称为 $U\cup V$ 中各元素之间(即本体和方法)的关系.

### 1.2 领域本体的语义相关性

将 DOKM 知识库中的领域模型记为 $DM$,本文提到的实体既可以是对象也可以是本体.

**定义 4(领域模型)**. 领域模型定义为一个加权无环图 $DM=(O,A,R)$.其中 $O$ 是图 $DM$ 的非空顶点集,由描述该领域知识模型的本体有限集构成;$A\in O$ 是图 $DM$ 的根节点,称为领域知识模型的根本体;$R\subseteq O\times O\times N$ 构成图的加权边,表示图中各领域本体的相关度(下文将给出定义),式中 $N$ 为自然数集合.

领域模型的根本体是区别领域模型的惟一标识,主要用于描述与整个领域模型相关的信息,将领域模型

$DM$ 的根本体记为 $DR(DM)$.$DM$ 图加权边的权 $w$ 给出了领域模型内各本体之间语义相关度的值,将领域本体 $O_1$ 与 $O_2$ 之间的语义相关度记为 $R^W(O_1,O_2)$.为了定义领域本体语义相关度,先引用文献[1]提出的非相关度的定义.

**定义 5(实体非相关度)**. 用 $ID(x,y)$表示 $x$ 与 $y$ 的非相关程度;用 $QID(x,y)$表示 $x$ 与 $y$ 的准非相关程度;用 $UB(x)$表示 $x$ 可取值范围的上界集.上述定义满足如下条件:

(1) $\max(ID(x,y),ID(y,z))\in UB(ID(x,z))$;

(2) 将式(1)中的 $ID(x,y)$用 $UB(ID(x,y))$替换,或/和将 $ID(y,z)$用 $UB(ID(y,z))$替换,式(1)仍成立;

(3) 对$\forall x\subset y$,有 $ID(x,y)=\min\{u|u\in UB(ID(x,z))\}$;

(4) $\max(QID(x,y),QID(y,z))\in UB(QID(x,z))$;

(5) 将式(4)中的 $QID(x,y)$用 $UB(QID(x,y))$替换,或/和将 $QID(y,z)$用 $UB(QID(y,z))$替换,式(4)仍成立;

(6) $QID(x,y)=\min\{u|u\in UB(QID(x,z))\}$;

(7) 仅在(1)~(6)描述的情况下,$ID(x,y)$具有有限值;其他情况下,$ID(x,y)=\infty$;

(8) 如果 $y$ 是 $x$ 的祖先,则 $ID(x,y)=0$.

两个元素间的非相关度越小,相互间的相关程度越大.我们规定:当 $ID(x,y)=0$ 时,$x$ 对 $y$ 是必需的.

**定义 6(领域本体语义相关度)**. 给定领域模型 $DM$ 内任意两个领域本体 $p$ 和 $q$,若 $p$ 与 $q$ 的非相关度为 $n$,$n\in N$ 且 $0\le n\le\infty$,则 $R^W_{DM}(p,q)$称为领域本体 $p$ 与 $q$ 的语义相关度,$0\le R^W_{DM}(p,q)\le 1$,且:

(1) $R^W_{DM}(p,q)=0$,当 $n=\infty$;

(2) $R^W_{DM}(p,q)=1$,当 $n=0$;

(3) $R^W_{DM}(p,q)=1/n$,当 $0<n<\infty$;

(4) $R^W_{DM}(p,q)=1$,当 $p>q$,即 $p$ 是 $q$ 的祖先节点;

(5) $R^W_{DM}(a,c)=\min(m,n)$,当 $R^W_{DM}(a,b)=n$, $R^W_{DM}(b,c)=m$.

**定义 7**. 给定领域模型 $D_1$ 和 $D_2$,称 $R^W_{D_1,D_2}$ 为两个领域模型根本体之间的语义相关度.DOKM 系统按如下方式确定 $R^W_{D_1,D_2}$ :(1) 如果直接由 DODL 文本提取领域本体,则由领域专家在设计 DODL 文本时确定描述元素的重要程度(〈必需〉、〈常用〉或〈任选〉),再经过 DOKM 系统编译处理转换为根本体的语义相关度.(2) 如果系统是由 VDO 构造系统根据现有领域本体生成新领域本体,则将新领域本体作为原领域本体的子本体,与原领域本体的语义相关度取值为 1.

**定义 8**. 给定领域模型 $D_1$ 中的领域本体 $p$ 和领域模型 $D_2$ 中的领域本体 $q$,则称 $R^w(p_{D_1},q_{D_2})$ 为领域本体 $p$ 和 $q$ 之间的语义相关度,如果有 $R^w(p_{D_1},q_{D_2})=R^W_{D_1}(p,DR(D_1))*R^W_{D_1,D_2}*R^W_{D_2}(DR(D_2),q)$.

## 2 虚拟领域本体构造方法

我们采用的仿生物种群进化方法包括选择(selection)、克隆(clone)、变异(mutation)、杂交(crossover)、合成(synthesis)和转基因(transgenic),用以构造虚拟领域本体.限于篇幅,以下仅讨论其中较典型的领域本体杂交、合成和转基因方法.

### 2.1 领域本体的杂交

领域本体杂交是指通过组合两个或两个以上的现有领域本体,构成新本体.两个领域本体杂交的情况可以表示为图 1,杂交后得到的目标领域本体结构如图 2 所示,进行杂交操作的两个初始领域本体依然被保留.杂交操作要求参与本体在属性、方法或关系方面有语义重叠(即语义相关度 $R^W_D(O_1,O_2)>0$).

**算法 1**. 给定领域本体 $O_1$,$O_2$ 以及所属领域模型 $DM_1$ 和 $DM_2$,求语义相关度为 $k$ 的新领域本体 $O$.

(1) 若 $O_1\not\subset DM_1$ 或 $O_2\not\subset DM_2$,则算法终止,否则继续;

(2) 若 $R^W_D(DR(DM_1),DR(DM_2))\ne m$ 且 $R^W_D(DR(DM_2),DR(DM_1))\ne n$($0<m,n\le 1$),则调用算法 2,否则继续;

(3) 分 3 种情况:

(i) 如果 $R^W_D(DR(DM_1),DR(DM_2))=m$ 并且 $R^W_D(DR(DM_2),DR(DM_1))\ne n$,则调用算法 3 在领域模型 $DM_2$ 内构造

以 $q$ 为基本体,$p$ 为扩展本体,语义相关度为 $m$ 的领域本体 $O$;

(ii) 如果 $R_D^W(DR(DM_1),DR(DM_2))\neq m$ 并且 $R_D^W(DR(DM_2),DR(DM_1))=n$,则调用算法 3 在领域模型 $DM_1$ 内构造以 $p$ 为基本体,$q$ 为扩展本体,语义相关度为 $n$ 的领域本体 $O$;

(iii) 如果 $R_D^W(DR(DM_1),DR(DM_2))=m$ 且 $R_D^W(DR(DM_2),DR(DM_1))=n$,则由用户选择领域模型 $DM_1$(或 $DM_2$),调用算法 3 在领域模型 $DM_1$(或 $DM_2$)内构造以 $p$(或 $q$)为基本体,$q$(或 $p$)为扩展本体,语义相关度为 $m$(或 $n$)的领域本体 $O$;否则:

(a) 若 $n<m$,则调用算法 3 在领域模型 $DM_1$ 内构造以 $p$ 为基本体(base-ontology),$q$ 为扩展本体,语义相关度为 $n$ 的领域本体 $O$;

(b) 若 $n>m$,则调用算法 3 在领域模型 $DM_2$ 内构造以 $q$ 为基本体,$p$ 为扩展本体,语义相关度为 $m$ 的领域本体 $O$;

(c) 若 $n=m$,则由用户选择 $DM_1$(或 $DM_2$),调用算法 3 在 $DM_1$(或 $DM_2$)内构造以 $p$(或 $q$)为基本体,$q$(或 $p$)为扩展本体,相关度为 $m$(或 $n$)的领域本体 $O$;

(4) 算法结束,领域本体 $O$ 为所求结果.

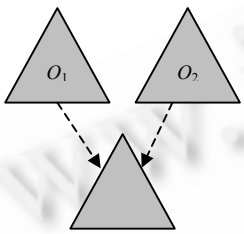


Fig.1 Crossover of domain ontology $O_1$ and $O_2$
图 1 领域本体 $O_1$ 和 $O_2$ 的杂交

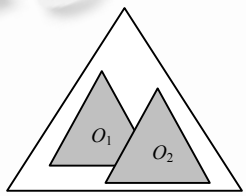


Fig.2 Target domain ontologies
图 2 目标领域本体

**算法 2**. 用户参与,以交互方式构造领域本体(算法描述略).

当两个领域本体所属领域模型不存在任何语义关联时,即 $R_D^W(DR(DM_1),DR(DM_2))=0$ 且 $R_D^W(DR(DM_2),DR(DM_1))=0$,则由用户使用交互方式选择或新输入本体的属性、方法及关系等.

**算法 3**. 在领域模型 $DM$ 内构造以 $p$ 为基本体($p\in DM$),$q$ 为扩展本体,语义相关度为 $m$ 的领域本体 $O$.

(1) 将基本体 $p$ 全部复制到领域本体 $O$ 中;

(2) 循环执行以下操作,直到 $DM$ 中再没有新领域本体 $h$,则结束循环.

(a) 若领域本体 $q\in DM$,对领域本体 $h\in DM$ 且 $R_D^W(q,h)=m$,则将 $h$ 构成部分复制到 $O$ 中;

(b) 若领域本体 $q\notin DM$,则求得 $q$ 所属领域模型 $DM'$,对 $\forall h\in DM'$,若 $R^w(O_{DM},h_{DM'})=m$,则将 $h$ 构成部分复制到 $O$ 中;

(3) 领域本体 $O$ 即为所求.

### 2.2 领域本体的合成

领域本体合成是指将领域知识库内符合要求的若干个领域本体进行适当组合,得到一个新本体的过程.与本体杂交不同的是领域本体合成需要给出一个初始本体,领域本体合成过程如图 3 所示.

领域本体合成可以直接在领域本体杂交算法的基础上建立起来,合成过程可以描述为:

给定领域基本体 $O_b$,领域模型 $DM$,语义相关度 $m$,在领域知识库 $DKB$ 中合成新本体 $O$,其中 $O_b\in DM$.

反复执行以下步骤(1)~(3),直至 $DKB$ 内不再含有新领域本体 $h$ 为止.

(1) 从 $DKB$ 中取出领域本体 $h\in DM'$;

(2) 对领域本体 $O_b$,$h$ 以及各自所属的领域模型 $DM$ 和 $DM'$,求本体语义相关度为 $k$ 的新领域本体 $p$;

(3) 将领域本体 $p$ 包含的各构成部分复制到 $O$ 中.最后即可得到领域本体 $O$.

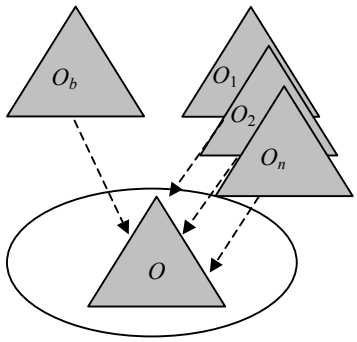


Fig.3 Construct a new ontology $O$ from given domain ontology $O_b$
图 3 以领域本体 $O_b$ 为基本体,构造新本体 $O$

### 2.3 领域本体的转基因

领域本体的转基因是指对已有的领域本体中的某些成分实施增删修改等操作.给定领域本体 $O$,领域本体 $P$,领域本体转基因操作可以区分为以下几种情况:

(1) 使用与 $P$ 语义相关的(包括 $P$)成分扩充领域本体 $O$;

(2) 从领域本体 $O$ 中删除与 $P$ 语义相关的(包括 $P$)领域本体的成分;

(3) 使用与 $P$ 语义相关的(包括 $P$)成分扩充领域本体 $O$,并删除领域本体 $O$ 中与 $P$ 语义相关的成分;

(4) 给定领域本体 $O$,用户直接对本体 $O$ 进行交互编辑.系统自动维护领域知识库的一致性和完整性.

## 3 VDO 构造系统设计与应用实例

虚拟领域本体 VDO 构造系统结构示意如图 4 所示.由用户(领域专家)输入虚拟领域本体构造需求,系统根据需求从领域知识库 DOKB 中选取符合要求的初始领域本体;然后继续由用户选择构造方式,如选择、克隆、变异、杂交、合成、转基因等,开始构造虚拟领域本体;在虚拟领域本体构造过程中,系统需要反复与用户交互,确认有关用户需求或领域描述知识;最后将构造好的虚拟领域本体存入领域本体知识库 DOKB 内.VDO 构造过程需要用户交互介入才能完成所有操作.目前,还不能在语义阶段实现虚拟领域本体的自动构造.

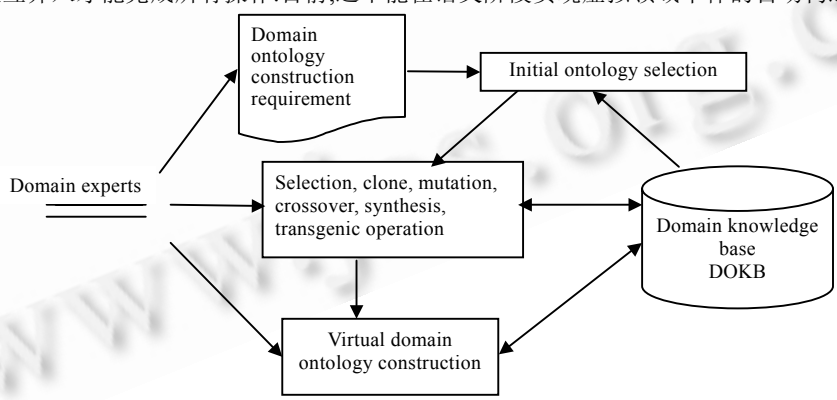


Fig.4 Architecture of VDO constructing system
图 4 VDO 构造系统

在 DOKM 领域知识库系统中,根据已有领域知识构造虚拟领域本体效果较好.下面给出一个实例:

设在 DOKB 知识库中已经存放有宾馆领域的一个标准三星级宾馆 $H$ 的领域模型表示,以及旅行服务领域某个跨国旅行服务社 $T$ 的领域模型表示.现在,为方便住客的旅游出行,增加客流量,宾馆 $H$ 决定在商务中心增设一个旅行代办机构 $A$.很显然,这需要对领域模型 $H$ 进行扩充,即在 $H$ 中新建一个虚拟领域本体 $A$,专门用于描述该旅行代办机构.限于篇幅,以下只给出使用 VDO 系统构造该领域本体的非形式化描述:

(1) 用户提出需要构造领域本体的请求;

(2) 用户选择新领域本体的领域模型、父类(如有的话)、属性描述、领域本体语义相关度.此处假设用户输入了三星级宾馆 $H$、商务中心、旅行社&代办机构&跨国旅行等;

(3) 系统采取匹配领域本体语义相关度以及概念类名的办法,选择初始领域本体;

(4) 这时用户有两种方式选择基本体 $O_b$.(a) 指定领域本体语义相关度 $K$,由系统选择适合的基领域本体;(b) 手工选择自己需要的基领域本体(可编辑);

(5) 系统运用前文叙述的领域本体合成算法,在三星级宾馆 $H$ 领域模型内,以商务中心领域本体为父本体,以 $O_b$ 为领域基本体,构造出新旅行代办机构本体 $A$.

## 4 相关工作

重用已有本体知识构造新领域本体的方法可以区分为两大类:一是从用户需求分析出发,在现有本体基础上构造新本体,现有本体将作为新本体的一部分出现;二是归并具有相同主题的现有本体中的概念、分类标准和规则,用以构造新本体.在第 1 类方法中,需要分析被选本体是否具有足够的描述细节和粒度,对于多本体库操作可能还要考虑到本体描述语言的转换等.Physical Systems 本体库是在 Ontolingua[4]上实现的一个本体重用系统,系统区分了 3 类本体集成方法:应用新的概念和关系对本体扩充、本体内容进行定制、分析现有领域本体提取领域间依赖关系构造新本体.该方法对领域依赖关系作了深入分析,但对如何利用这些关系构造新本体有待进一步讨论.另外,利用现有自然语言本体(特别是 MRDs 机器可读字典)作为本体构造资源的研究也很多,如:在 Ontosaurus[5]中利用了 SENSUS[6]作为 MRDs,采用半自动化的办法构造领域本体,DODDL[7]则以 WordNet[8]作为 MRDs 对现有的领域本体作有限的修剪操作.第 2 类方法主要是针对具有相同主题的多个本体,作归并意义上的集成,以构造新的本体.Gertjst van Heijst 从改造现有的同类医药领域本体库出发,对医药本体库中本体的概念类属性以及关系进行了整合和适当扩充[9].我们提出的虚拟领域本体构造方法与以上方法不同,我们是采用本体语义相关度匹配的办法来搜索和匹配本体,增大了系统的适用范围.

## 5 结束语

本文主要讨论了如何重用现有领域知识库中的知识,构造新领域本体的方法.在语义级实现本体自动归并、合成操作是极不现实的,我们采用了一种手工交互操作与系统自动分析相结合的办法,该方法在一定程度上增强了系统的实用性.我们下一步的主要计划是从本体的深层次语义分析入手,分析本体的重用问题.

**References**:

[1] Lu RQ, Jin Z. Domain Modeling Based Software Engineering. Kluwer Academic Publishers, 2000. 1~347.

[2] Lu RQ, Jin Z. Ontology-Oriented requirement analysis. Journal of Software, 2000,11(8):1009~1017 (in Chinese with English Abstract).

[3] Jin Z. Ontology-Based requirements elicitation. Chinese Journal of Computers, 2000,23(5):486~492 (in Chinese with English Abstract).

[4] Gruber TR. A translation approach to portable ontology specification. Knowledge Acquisition, 1993,5:199~220.

[5] Swartout B, Patil R, Knight K. Toward distributed use of large-scale ontologies. In: Proceedings of the 10th Knowledge Acquisition Workshop (KAW'96). 1996. http://ksi.cpsc.ucalgary.ca/KAW/KAW96/swartout/Banff_96_final_2.html.

[6] Knight K, Luk S. Building a large knowledge base for machine translation. In: Proceedings of the AAAI'94. 1994.

[7] Yamaguch T. Constructing domain ontologies based on concept drift analysis. In: Proceedings of the IJCAI'99 Workshop on Ontologies and Problem-Solving Methods (KRR5). 1999. http://sunsite.informatik.rwth-aachen.de/Publiccations/CEUR-WS/Vol-18/.

[8] Miller G. WordNet: an on-line lexical database. International Journal of Lexicographer, 1990,3(4):234~244.

[9] Heijst G. The role of ontology in knowledge engineering [Ph.D. Thesis]. Amsterdam: University of Amsterdam, 1995.

**附中文参考文献:**

[2] 陆汝钤,金芝,陈刚.面向本体的需求分析.软件学报,2000,11(8):1009~1017.

[3] 金芝.基于本体的自动需求获取.计算机学报,2000,23(5):486~492.