一种基于概念格的本体合并方法

盛艳，李云，李拓，袁运浩

（扬州大学 信息工程学院，江苏扬州 225009）

摘　要：本体合并是指将相同或者相似领域内已存在的本体合并在一起，消除重叠的和不协调的部分，从而实现现有知识系统的共享、重用及扩展。目前，已有很多研究者提出了各种本体合并的方法，其中，利用概念格进行本体合并的方法非常有效。提出了一种基于概念格的，并利用叙词表的方法进行本体合并。为了进一步获得提取本体概念的相关指导，提高本体概念抽取的自动化程度，提出最小外延集概念，从而更方便有效地进行本体合并。

关键词：本体合并；概念格；叙词表；最小外延集

中图分类号：TP18　　**文献标识码**：A　　**文章编号**：1000-7180 (2008) xx-xxxx-x

A method for ontology merging based on the concept lattice

SHENG Yan, LI Yun, LI Tuo, YUAN Yun-hao

(The Information Engineering College of Yangzhou University, Yangzhou Jiangsu 225009)

Abstract: Ontology merging is to amalgamate two or more existing ontologies in the same or similar areas to eliminate their overlapping and uncoordinated parts, thus achieving the sharing, reusing and expansion of knowledge systems. At present, many researchers put forward lots of ways for ontology merging, which using concept lattice is very effective. This paper advanced a method using thesaurus for ontology merging based on the concept lattice model. In order to extract the ontology concepts more conveniently and automatically, the notion SLET is advanced, thereby we can carry out ontology merging effectively.

Key words: ontology merging; concept lattice; thesaurus; SLET

1. 引言

本体合并是指将相同或者相似领域内已存在的本体合并在一起，消除重叠的和不协调的部分。近些年来有学者提出支持合并本体的一些系统和框架，例如 Onto-Morph 系统[1]、Chimaera 系统[2]、Protēgē2000 中的 Prompt 算法[3]等等，大部分依赖于句法和语义的启发式方法，没有提供对整个合并过程的全局性描述。本文提出了一种基于概念格，并借助叙词表的方法进行本体合并。

2. 背景知识

概念格是形式概念分析(FCA) [4] 的核心数据结构，反映对象和属性之间的联系以及概念间泛化和例化关系。形式背景被定义为一个三元组 K=(G, M, I)，其中 G 和 M 分别是对象集合和特征集合，而 I 是 G 和 M 之间的二元关系，即 I⊆G×M，gIm。

在形式背景 K 中，在 G 的幂集和 M 的幂集之间可以定义如下的两个映射函数 f 和 g：

∀A⊆G: f(A)={m|∀x∈A(xIm)}

∀B⊆M: g(B)={g|∀y∈B(gIy)}

来自 P(G) ×P(M)的二元组(A, B)若满足两个条件：A=g(B)和 B=f(A)，则称(A, B)为形式背景 K 的一个形式概念，记为 C=(A, B)， B 和 A 分别被称为 C 的内涵(Intent)和外延(Extent)。若 C1=(A1，B1)和 C2=(A2，B2)，满足 A1⊆A2，则称(A1，B1)为亚概念，(A2，B2)为超概念，记为：(A1，B1) ≤ (A2，B2)。这种超概念－亚概念的偏序关系所诱导出的格称为概念格。

本体最初是一个哲学上的概念，用于描述事物的本质。Gruber[5]的定义被引用最多的：本体是共享概念化的形式化、显式的说明。许多定义共享了关于本体最核心的部分：本体概念以及本体概念之间的 is-a 关系。在概念格中，概念则存在一种偏序关系，从某种程度上来说，本体概念间的 is-a 层次关系和概念格中的偏序关系非常相似，因此近年来 FCA 在本体中的应用很受到人们重视。

叙词表[6]（T，≥）是一系列具有偏序关系的项 ti∈T 的集合，它以树的方式呈现出来。每一颗树都

收稿日期：

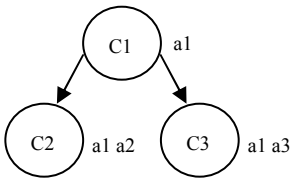
基金项目：国家自然科学基金(60575035，60673060)。

包含了项的定义以及该项所相关的父子项。

定义 1：叙词表（T，α）是一系列具有相似关系的项 ti∈T 的集合，它以树的方式表现出来，树中包含了项与项之间的层次关系。

3. 基于概念格的本体合并思想

在本体 O 如图 1 所示，C1 拥有属性 a1，C2 和 C3 分别是 C1 的子概念，则 C2 和 C3 必然继承属性 a1，且 C2，C3 自身独有的属性 a2，a3 在叙词表中必然处在属性 a1 的下一层。以本体概念作为形式背景中的对象，以概念所拥有的属性作为形式背景中的属性构建概念格如图 2，则由形式概念分析的理论可知 C2，C3 必然作为 C1 的子概念存在于 Hasse 图中。



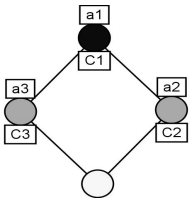


图 1 本体 0　　图 2 本体 0 对应的概念格

基于叙词表上述的特点，本体合并过程可分为以下三步：

（1）输入本体，构建形式背景。输入信息包括两个本体和叙词表，得到的形式背景如图 3 所示，叙词表如图 4 所示。

Default ...	att_0	att_1	att_2	att_3	att_4	att_5	att_6	att_7	att_8	att_9
obj_0	X	0	0	X	0	0	X	0	X	X
obj_1	0	X	X	0	0	X	X	0	0	0
obj_2	0	X	X	0	X	0	0	0	X	X
obj_3	X	0	0	X	X	0	X	0	X	0
obj_4	0	0	X	0	0	0	0	X	X	X
obj_5	0	X	0	0	0	X	0	0	X	X
obj_6	0	0	0	X	X	X	X	X	0	0
obj_7	X	0	0	X	0	0	X	X	X	X

图 3 根据输入信息构建的形式背景

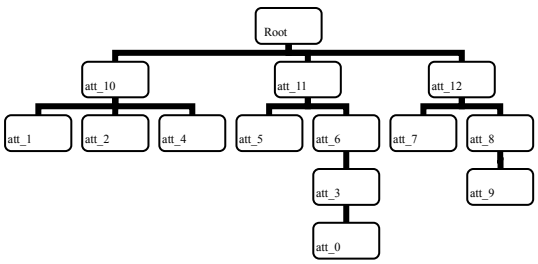


图 4 叙词表结构

其中 obj_0~obj_3 表示本体 O1 中的概念，obj_4~obj_7 表示本体 O2 中的概念，att_0 等表示利用叙词表得到的本体属性。拥有属性 att_0 的对象必然拥有属性 att_3，也必然拥有属性 att_6。

（2） 构造概念格。采用经典的渐进式构格算法得到如图 5 所示。

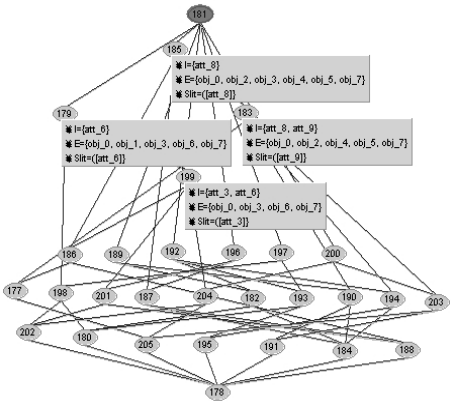


图 5 对应的概念格图

（3）分析概念格中有效节点，获得合并后的本体概念。

定义 2：在一个同属性的对象集集合中有一个对象集 X，若不存在另一个对象集 X'，满足 X'⊂X，则称 X 为该对象集中的最小对象集。所有最小对象集构成的集合简称为最小外延集(the Set of the Least ExTents)，简记为 SLET。给定形式概念(A, B)，则 SLET {s}满足以下两个条件：（1）f(s)=f(A)=B；（2）对任意 s'⊂s，f(s)⊂f(s')，即 SLET 同样是对格节点中的冗余信息进行删减的结果，能以最少的信息精确地反映格节点的本质。

分析顶层的四个节点，分别提取他们的 SLET，对 SLET 的势按照以下原则进行分析：

①如果格节点有一个势为 1 的 SLET：此时可以获得唯一的本体概念。如 199#节点的 SLET 为{{obj_6}}，因此需要提取 obj_6 作为合并后的本体概念名称。

②如果格节点有多个势为 1 的 SLET：这种情况意味着属性过少，无法区分 SLET 中的每个元素，需要领域专家考虑对叙词表进行处理，考虑 SLET 中元素的不同，重新添加属性，构造概念格。如 179#节点的 SLET 为{{obj_0}，{obj_7}，{obj_1，obj_3}}，此时领域专家应综合考虑本体概念 obj_0 和 obj_7 的不同，然后对叙词表进一步处理，完善形式背景属性集后，再次提取概念。

③如果格节点有势大于 1 的 SLET：在这种情况下，系统应提示本体开发者此处可能会产生新的概念，并且由本体开发者考虑是否需要获得新的本体概念。如 185#节点的 SLET 为{{obj_4，obj_5}，{obj_3，obj_4}，{obj_3，obj_5}，{obj_2，obj_7}，{obj_5，obj_7}，{obj_0，obj_2}，{obj_0，obj_4}，

{obj_0，obj_5}}。此时应领域专家应根据领域知识，提取一个包含 185#节点所有内涵的本体概念作为合并后的本体概念名称。对 183#节点的分析同 185#。

利用三条规则对概念格分析后，可以容易的提取出四个合并后的本体，而本体概念之间的层次关系则清晰的反映在 Hasse 图中。

4. 实验分析

为了检验利用 FCA 方法所合并的最终本体的质量，采用 Maedche[8]所提出的“黄金标准”作为衡量准则，参考本体使用 Australian Sustainable Tourism Ontology（AuSTO），叙词表采用了 Thesaurus on Tourism and Leisure Activities(TTLA)。

经过对 Yahoo 旅游项目信息中的网页进行分析后，提取了三组源本体，指定最小支持度为 20%，指定合并层次为叙词表 TTLA 的第四层次，利用上述方法对源本体进行两两合并后得到 O1，O2，O3。最后采用 AuSTO 本体作为标准本体与 O1，O2，O3 进行比较，利用黄金标准对结果进行衡量后最终得到的结果如表 1。

表 1 实验结果

	$\overline{TO}_{(\%)}$	L O (%)	F (%)
本体 01	27.56	27.28	27.42
本体 02	27.47	26.32	26.88
本体 03	27.19	28.95	28.04

对同样的数据集，采用 FCA-Merge 方法，指定最小支持度为 20%，最终得到的结果如表 2

表 2 FCA-Merge 实验结果

	$\overline{TO}_{(\%)}$	L O (%)	F (%)
本体 01	28.11	22.17	24.78
本体 02	26.62	21.49	23.78
本体 03	27.28	22.14	24.44

从结果中可以看出 FCA-Merge 方法得到结果的 LO 值显著低于使用叙词表后的 LO 值，而使用叙词表并未对$\overline{TO}$值有显著影响，因此利用自动化方法 FCA 来合并本体，使用叙词表后得到的最终结果是令人满意的。

5. 结束语

本文利用概念格与叙词表为本体合并提供了另外一种方法。利用叙词表统一了本体概念属性的表示方法；根据 SLET 的特点对概念格中的概念进行了分析，获得合并后的本体概念，并由 Hasse 图得到了合并后的本体概念之间的关系。实验表明本文的方法是有效的。

参考文献:

[1] CHALUPSKY H. OntoMorph : A translation system for symbolic knowledge . Proc. KR'00, Breckenridge, CO, USA, 471-482.

[2] McGUINNESS D L, FIKES R, RICE J, et al. An environment for merging and testing large ontologies. Proc. KR'00, 483-493.

[3] FRIDMAN NO Y N, MUSEN M A. PROMPT. Algorithm and tool for automated ontology merging and alignment. Proc. AAAI'00, 450-455.

[4] B.Ganter, R.Wille, Formal Concept Analysis:Mathematical Foundations, Springer, Heidelberg. 1999.

[5] T.R. Gruber, A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition ,5(2), 1993:199-221.

[6] de Souza, K.X.S., Davis, J.: Aligning ontologies through formal concept analysis. Journal on Data Semantics V, LNCS 3870, 2006：211–236,

[7] 李云，刘宗田，陈崚，蔡俊杰，量化概念格及其渐进式构造，模式识别与人工智能，Vol.19,No.3,2006:375-381

[8] A. Maedche and S. Staab. Measuring similarity between ontologies. In Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW). Springer, 2002：251-163

作者简介：

盛艳(1985-)，女，江苏苏州人，硕士研究生，研究兴趣为概念格，信息检索等；

李云(1965-)，男，安徽合肥人，博士，教授，研究兴趣为概念格，数据挖掘等；

李拓(1983-)，男，硕士研究生，研究兴趣为概念格，数据挖掘等；

袁运浩(1983-)，男，江苏徐州人，硕士研究生，研究兴趣为概念格，数据挖掘等。