基于概念格模型的本体映射*

盛艳，李云，李拓，栾鸾

（扬州大学 信息工程学院，江苏扬州 225009）
E-mail: shengyan1985@gmail.com

摘 要：本体映射是实现本体复用性的一个途径，目前已经存在一系列的方法解决这个问题。但现存的方法没有对本体进行前期处理，所提取的关系比较单一，忽略了本体概念所具有的层次关系的重要特点。本文利用形式概念分析对现有的本体映射方法进行了改进，首先利用信息熵对属性语义相似表进行定义，进而利用它统一了本体概念属性的表示方法，然后提出新的算法完善形式背景，利用完善后的形式背景对本体概念之间的相似度进行衡量，并通过概念格提取了除已知关系之外的多种新关系。

关键词：本体映射；形式概念分析；信息熵；语义相似度量

中图分类号：TP18　　**文献标识码**：A

Ontology mapping based on the lattice model

Shengyan, Li yun, Li tuo, Luanluan

(The Information Engineering College of Yangzhou University, yangzhou jiangsu 225009)
E-mail: shengyan1985@gmail.com

Abstract: Ontology mapping could achieve the reuse of the ontologies, and there are lots of methods for ontology mapping. However, few methods think much of the concepts' hierarchy and precondition the ontologies, so the kinds of the relations extracted are very simple. In this paper we improve the existing method by using formal concept analysis. First, semantic similarity matrix using information entropy is given in order to unify the description of the ontology concepts' attributes. Second, a new algorithm is advanced to complete the formal context. Third, the similarity of the ontology concepts is computed by the completed formal context and new relations could be extracted besides others in the existing literature.

Keywords: ontology mappings; formal concept analysis; information entropy; semantic similarity

1. 引言

随着语义网的不断发展，由不同组织开发所得的本体数量随之增加，因此将相同或者相近领域内的本体进行映射是很有必要的。目前已经存在多种本体映射方法，比如 Pan[1], Stoilos[2], Euzenat[3], Castano[4]等都提出了不同的映射方法，但这些方法只能获取本体概念之间的等价关系，而对于一些其他关系（如层次关系等）没有被提取出来。

利用概念格表示本体模型并且进行本体相似度的衡量是非常有效的。Fan[5]提出了一种利用形式概念分析进行本体相似度衡量的方法，除了提取出本体概念之间的等价关系以外，进一步提取了本体概念之间的层次关系。但是由于本体表现方式的多样性，利用格结构所提取的两种关系类型依然是不全面的。本文基于概念格与属性语义相似表提供了另外一种本体映射方法，并提取了本体概念之间的多种关系，弥补了以上诸多方法的不足。

2. 背景知识

2.1 形式概念分析

*基金项目：国家自然科学基金(60575035，60673060)。

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；栾鸾(1985-) ，女，江苏扬州人，硕士研究生，研究兴趣为概念格，数据挖掘等。

形式概念分析 Formal Concept Analysis(FCA) [6]是一种建立在数论基础上的数据分析技术，其基本内容是形式背景(Formal Context)、形式概念(Formal Concept)以及概念之间的关系。形式背景(Formal Context)被定义为一个三元组 K=(G, M, I)，其中 G 和 M 分别是对象(object)集合和特征(属性 attribute)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)}

它们也称为 U 和 A 之间的 Galois 连接。来自 P(G)×P(M)的二元组(A, B)如果满足两个条件：A=g(B)和 B=f(A)或 A=B’和 B=A’，则称(A, B)为形式背景 K 的一个形式概念，记为 C=(A, B)，其中 B 和 A 分别被称为形式概念 C 的内涵(Intent)和外延(Extent)。设 C1=(A1,B1)和 C2=(A2,B2)是两个形式概念，其中偏序关系“≤”定义为 C1≤C2⇔B2≤B1。此时称 C1 是 C2 的子概念,C2 是 C1 的超概念。这样，由形式概念及父子关系就可形成一个概念格。

2.2 本体

本体被引入到计算机领域，尤其是人工智能及其相关领域中，主要用于知识表达以及知识共享。随着人们对本体理解的不断加深，提出本体最核心的部分：本体概念以及本体概念之间的 IS-A 关系。本体的描述模型主要分为基于框架的描述模型和基于逻辑的描述模型。本体概念中存在被称为 IS-A 的层次关系，而概念格中的形式概念则存在一种偏序关系，从某种程度上来说，二者非常相似，因此近年来 FCA 在本体中的应用受到人们重视。FCA 除了可以自动的构建本体之外，还可以利用概念格中格节点的特性来衡量本体之间的相似度，而本体映射关系就是以本体相似度衡量为基础的。

2.3 信息熵

信息熵最主要的思想是：如果有一个系统 S 内存在多个事件 S={E1，E2,…,En}，每个事件的概率分布 P={P1，P2，…，Pn}，则每个事件本身的信息量为 $I(E)=-\log Pi$ (即当事件 Ei 出现后，给予我们的信息量)，整个系统的平均信息量为 $H(S)=\sum_{i=1}^{n} PiI(E)=-\sum_{i=1}^{n} Pi\log Pi$。这个平均信息量就是信息熵，简称熵。通常规定如果 Pi=0，则 I(E)=0，这与数学上极限定理的定义是一致的。

本文中主要是用来计算每个属性的信息量，以此来计算属性间的语义相似度量。

3 基于概念格的本体映射及其关系提取

3.1 属性间语义相似度量

为了能有效地从语义上衡量属性之间的相似度，采用 WordNet[7]作为基础词汇信息库，提取词汇间的 IS-A 关系。然后在具体领域的文集中，统计各个词汇出现的概率，得到一个加权的 IS-A 层次关系。接着通过计算各个词汇的信息量得到语义相似度，从而得到词汇语义相似表。

WordNet 是一个覆盖范围宽广的英语词汇语义网，本文中主要使用 WordNet 中词汇之间的 IS-A 关系和同义词集 SynSet。对于具体领域的文集，统计各个词汇出现的概率，并加入 IS-A 层次关系中，得到一个加权的 IS-A 层次关系如图 1 所示。

定义 1 加权 IS-A 层次关系：给定一个英文词汇库 $\mathcal{E}$，利用每个词汇概率及其间的 IS-A 关系构成加权 IS-A 层次关系 $H(\varepsilon)$。其中词汇概率定义为：$p(n)=\frac{freq(n)}{M}+\sum_{i\in sub(n)} p(i)$，

其中，$freq(n)$ 为词汇 n 在文集中出现的次数；M 为文集中所有词汇的数目；$\sum_{i\in sub(n)} p(i)$ 为词汇 n 的所有直接下层词汇的概率之和，若词汇 n 是底层节点，即没有下层节点，则该值为 0。同时，定义一个层次关系的最顶层的节点，记为 Top，其 $p(Top)=1$。

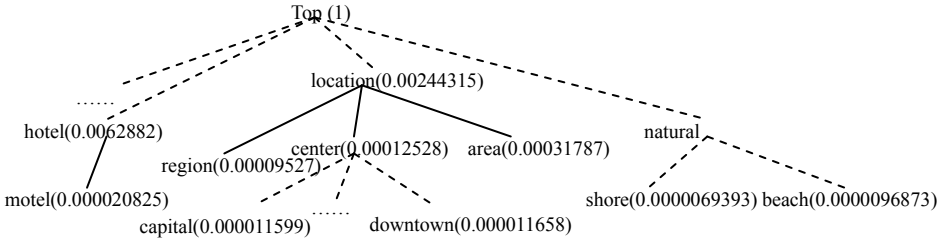

图 1 加权 IS-A 层次关系

从上述加权 IS-A 层次关系图中，根据信息量的计算公式 $I(E)=-\log Pi$，得到各个词汇的信息量，比如：

$I(downtown)=-\log 0.000011658=4.933376$，$I(area)=-\log 0.0031787=2.497751$。

可以看到在加权 IS-A 层次关系图中越上层的词汇所包含的信息量越少。并且对于两个词汇，若它们的共同父节点词汇的信息量越大，就表示它们共享的信息就越多，从而表示它们相似度越大。本文采用文献[8]中提出了信息量相似度作为词汇相似度量，以此进一步衡量属性间的语义相似度。

定义 2 信息量相似度 ics(n_1,n_2)：给定一个英文词汇库 ε 及加权 IS-A 层次关系 $H(\varepsilon)$，对于文集中的任意两个词汇 n_1，n_2，若 n_1= n_2或者 n_1和 n_2是同义词，则 $ics(n_1,n_2)=1$；否则 $ics(n_1,n_2)=\dfrac{2I(n')}{I(n_1)+I(n_2)}$；其中，n’ 为 n_1、n_2的最大公共父节点词汇，即 $I(n')=\max\limits_{n\in S(n_1,n_2)}\{I(n)\}$，$S(n_1,n_2)$ 是所有 n_1、n_2的公共父节点词汇。

那么，根据这个定义，可以得到任意两个词汇间的相似度。比如：

$ics(shore,beach)=1$

$$ics(capital,downtown)=\frac{2I(center)}{I(capital)+I(downtown)}=\frac{2*3.902118}{4.935579+4.933376}=0.790786\,3$$

.2 构建形式背景并生成概念格

假设有如下两个本体 O_1和 O_2如图 2 所示。

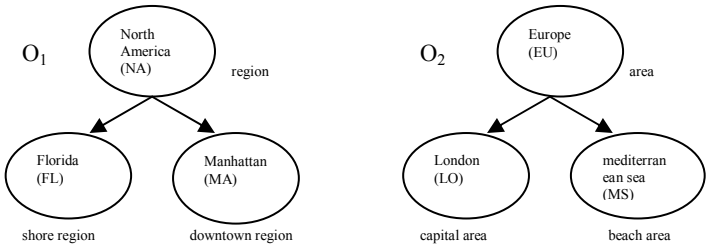

图 2 本体 O_1和本体 O_2

提取两个本体 O_1和 O_2的属性，根据词汇间的信息量相似度得到属性语义相似表表 1。

表 1 属性语义相似表

ics	area	capital	beach
region	1	0.583397	0
shore	0	0	1
downtown	0.703002	0.790786	0

不同的属性隶属于不同的本体，为了将不同的本体结合起来，应当构建统一的形式背景。构建形式背景的步骤如下：

（1）选择需要映射的本体概念(即作为形式背景中的对象)，并且获得它们的属性(即作

为形式背景中的属性)。

（2）对于本体概念原有属性，在形式背景中相应置为 1。但由于不同本体间的属性之间存在相似性，则可以利用属性语义相似表完善形式背景，提出算法 1 来完善形式背景。

算法 1：基于属性语义相似度量的形式背景的完善算法
PROCEDURE Complete_Context(O_1, O_2，The_Similarity_Matrix)
BEGIN
  FOR 每个 $a_{1i}\in A1$，$a_{2j}\in A2$，$i\in\{1,...p\}$，$j\in\{1,...q\}$
    根据属性语义相似表得到 ics (a_{1i}, a_{2j})
  END FOR
  FOR 每个概念 $C_{1s}\in O_1$
    获取 C_{1s} 拥有的所有属性在相似矩阵中组成的矩阵
    获得每一列的最大值 MAX_{2v}, $v\in\{1,...q\}$
    C_{1s} 行中 a_{2v} 的值置为 MAX_{2v}，$v\in\{1,...q\}$
  END FOR
  FOR 每个概念 $C_{2t}\in O_2$
    获取 C_{2t} 拥有的所有属性在相似矩阵中组成的矩阵
    获得每一行的最大值 MAX_{1u}, $u\in\{1,...p\}$
    C_{2t} 行中 a_{1u} 的值置为 MAX_{1u}，$u\in\{1,...p\}$
  END FOR
END

比如对于概念 MA，其拥有 region，downtown 两个属性，根据相似矩阵中的词汇 area，capital，beach 分别比较之后，分别取得 region，downtown 两行里较大的相似度值，即 ics(region, area)=1, ics(downtown, capital)=0.790786, ics(are, beach)=0。

根据算法 1 获得的初步形式背景如表 2

表 2 初步形式背景

	region	shore	downtown	area	capital	beach
NA	1			1	0.583397	0
FL	1	1		1	0.583397	1
MA	1		1	1	0.790786	0
EU	1	0	0.703002	1		
LO	1	0	0.790786	1	1	
MS	1	1	0.703002	1		1

对于已经构造完成的形式背景，应通过设定阈值过滤掉一些相似度较低的数据，从而使形式背景的内容更为精确，此处取阈值为 0.7，然后用×代替形式背景中的值，得到最终的二值化形式背景如表 3，利用 Godin[9]算法可以构建完整的概念格如图 3。

表 3 二值化的形式背景

	region	shore	downtown	area	capital	beach
NA	×			×		
FL	×	×		×		×
MA	×		×	×	×	
EU	×		×	×		
LO	×		×	×	×	
MS	×	×	×	×		×

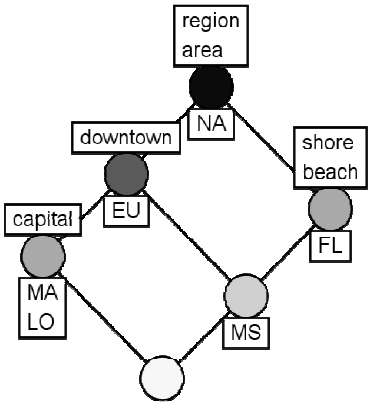

图 3 由表 3 生成的概念格

3.3 概念相似度量

本文中采用文献[8]中的类似方法，但概念的本质是通过概念所拥有的属性来表现的，所以只需要对概念所拥有的属性的不同进行衡量，即在这里只考虑形式概念的内涵。

定义 3 Sim(C1，C2)：两形式概念 $C_1(A_1, B_1)$ 和 $C_2(A_2, B_2)$ 的相似度 $Sim(C_1,C_2)=\frac{M(B_1,B_2)}{\max\{|B_1|,|B_2|\}}$，其中，$M(B_1,B_2)=\max\{\sum_{b_1\in B_1,b_2\in B_2} ics(b_1,b_2)\}$，$b_1$ 和 b_2 只能在每一种组合中出现一次。

比如考虑两个形式概念 C_1((MA,LO),(regin,capital,area,downtown)) 和 C_2((FL),(region,shore,area,beach))：

$$Sim(C_1,C_2)=\frac{M((region,capital,area,downtown),(region,shore,area,beach))}{\max\{|(region,capital,area,downtown)|,|(region,shore,area,beach)|\}}=0.5$$

3.4 多种关系的提取

通过对图3的分析,可以很容易的获得概念之间的各种关系,Fan[5]提取了以下两种关系：

（1）相等关系(Equal)，即概念 C_1 和 C_2 的属性完全一样，Equal(C_1，C_2)=Sim(C_1, C_2)=1。图 3 中，存在相等关系的有：Equal(MA,LO)。

（2）父子关系(Sub)，即概念 C_1 和 C_2 的属性具有包含关系，Sub(C_1，C_2)= Sim(C_1, C_2),这个相似度应该是很大的。图 3 中，存在父子关系的有：Sub(NA,FL)，Sub(EU,MS)，Sub(EU,LO)。

再考虑概念 MS 和 MA 两个节点，他们具有除根节点以外的共同父节点，这就说明 MA 和 MS 存在相同的属性，即是存在某种关系的。

定义 4：如果概念格中两个不同概念 C_1，C_2 之间不存在父子关系，并且存在除根节点以外的共同父概念，那么这两个概念之间存在的关系称之为交迭关系，记为 Overlap(C_1,C_2)。其间相似度为 Sim(C_1,C_2)。

根据定义 4，发现存在交迭关系的有: Overlap(MS,MA)， Overlap(MS,LO)。

定义 5：如果概念格中两个不同概念 C_1，C_2 之间不存在父子关系，也不存在交迭关系，但是他们的语义相似度 Sim(C_1,C_2) 大于领域专家设定的一个阈值 λ，那么这两个概念之间存在的关系称之为相似关系，记为 Similarity(C_1,C_2)。

根据定义 5，设定 $\lambda=0.5$，得到以下节点存在相似关系：Similarity(FL,MA)，Similarity(FL,LO)。

定义 6：如果概念格中两个不同概念 C_1，C_2 之间都不存在以上各种关系，那么，我们定义概念 C_1 和 C_2 不存在任何关系，记为 None(C_1,C_2)=0。

通过提取以上概念间的各种关系，丰富了概念间的映射关系，同时，我们也可以很清楚的看到，各种关系的相似度值有以下规律：Equal>Sub>Overlap>Similarity>None。

4 实验分析

本文首先抓取 Yahoo 旅游项目[10]中的网页，使用本体生成工具 Text2Onto[11]提取二组源本体。根据 WordNet2.0[7]产生加权 IS-A 层次关系，从而得到属性语义相似表。然后提取了相关的本体概念，并且利用属性语义相似表中的词汇表示本体概念属性，生成形式背景后使用 ToscanaJ-1.6[12]生成概念格，接着分析概念格中形式概念及其关系提取本体概念的各种映射关系。最后采用了 Ehrig[13]的方法对映射的结果进行了评估：

$$r=\frac{\text{正确找到的映射}}{\text{可能存在的映射}},\quad p=\frac{\text{正确找到的映射}}{\text{所有找到的映射}},\quad f=2pr/(p+r)$$

对于第一组本体，我们获得了 45 个属性，得到 r=0.882，p=0.858，f=0.869；对于第二组本体，我们获得了 58 个属性，得到 r=0.891，p=0.885，f=0.887。图 4 展示了对两组本体进行映射后获得的 r, p, f 的值，可以看出采用本文提供的方法进行本体映射可以获得较好的结果。与其他方法相比，利用形式概念分析可以获得的关系类型更为丰富。

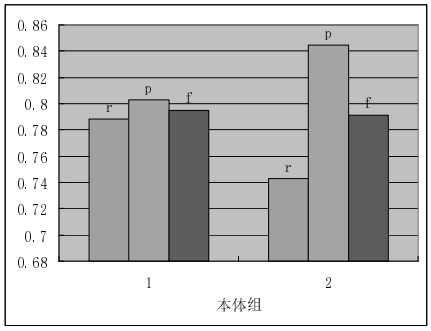

图 4 评估结果

5 结束语

本文利用形式概念分析结合属性语义相似表对现有的本体映射方法进行了改进，首先利用属性语义相似表统一了本体概念属性的表示和相似方法；然后利用算法对形式背景做了进一步处理，使其保留的信息更加完整；最后采用概念内涵的相似度来衡量了本体概念之间的相似度，并且通过概念格的结构获得了本体概念之间不同类型的关系。实验表明本文的方法是有效的。

参考文献

[1] Pan, L.Y., Song, H., Ma, F.Y.. A Macrocommittees Method of Combining Multistrategy Classifiers for Heterogeneous Ontology Matching. In: Li, Q., Wang, G., Feng, L. (eds.): Advances inWeb-Age Information Management. WAIM 2004. Lecture Notes in Computer Science, Vol. 3129. Springer-Verlag, Berlin Heidelberg New York ,2004 :672-677

[2] Stoilos, G., Stamou, G., Kollias, S.. A String Metric for Ontology Alignment. In: Gil, Y., Motta, E., Benjamins, V.R. et al. (eds.): The Semantic Web – ISWC 2005. LNCS 3729. Springer-Verlag, Berlin Heidelberg New York ,2005: 624-637

[3] Euzenat, J., Valtchev, P.. Similarity-based ontology alignment in owl-lite. In: Proceedings of the European Conference on Artificial Intelligence. 2004: 333-337

[4] Castano, S., Ferrara, A., Montanelli, S.. Matching Ontologies in Open Networked Systems: Techniques and Applications. In Spaccapietra, S., et al. eds. Journal on Data Semantics V, LNCS 3870. Springer-Verlag, Berlin Heidelberg New York ,2006: 25-63

[5] Liya Fan, Tianyuan Xiao. FCA-Mapping: A Method for Ontology Mapping. 4th European Semantic Web Conference 2007

[6] B.Ganter , R.Wille. Formal Concept Analysis: Mathematical Foundations. Springer-Verlag, 1999.

[7] ftp://ftp.cogsci.princeton.edu/pub/wordnet/2.0/WordNet-2.0.exe

[8] Anna Formica. Concept similarity in Formal Concept Analysis: An information content approach. Knowledge-Based Systems archive Volume 21, Issue 1 (February 2008)

[9] Godinr, Missaouir, Alaouih. Incremental concept formation algorithms based on galois (concept) lattices[J]. Computational Intelligence, 1995, 11 (2) : 246-267.

[10] http://travel.yahoo.com/

[11] http://ontoware.org/projects/text2onto/

[12] http://toscanaj.sourceforge.net/

[13] Ehrig M., Staab S.. QOM - quick ontology mapping. In: Harmelen, F.V., McIlraith, S., Plexousakis, D. (eds.): The Semantic Web – ISWC 2004. LNCS 3298 . Springer-Verlag, Berlin Heidelberg New York ,2004 :683-696