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)

[摘要] 利用形式概念分析对现有的本体映射方法进行改进, 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] TP 18 [文献标识码] A [文章编号] 1672-1292(2008)04-0091-04

Ontology Mapping Based on the Concept Lattice Model

Sheng Yan, Li Yun, Li Tuo, Luan Luan

(Institute of Information Engineering, Yangzhou University, Yangzhou 225009, China)

Abstract In this paper, the existing method by using formal concept analysis is improved. Firstly, semantic similarity matrix using information entropy is given in order to unify the description of the ontology concepts' attributes. Secondly, a new algorithm is suggested to complete the formal context. Thirdly, the similarity of the ontology concepts is computed by the completed formal context and new relations other than the known relations could be extracted through the concept lattice.

Key words ontology mappings; formal concept analysis; information entropy; semantic similarity measure

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, 而一些其它关系(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.

1 基于概念格的本体映射及其关系提取

1.1 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

收稿日期: 2008-06-18

基金项目: 国家自然科学基金(60575035, 60673060和 60773103)、江苏省自然科学基金(BK2008206)和江苏省教育厅自然科学基金(08KJB520012)资助项目.

通讯联系人: 盛 艳, 硕士研究生, 研究方向: 概念格和信息检索等. E-mail: shengyan1985@gmail.com

个加权的 IS－A层次关系如图 1所示.

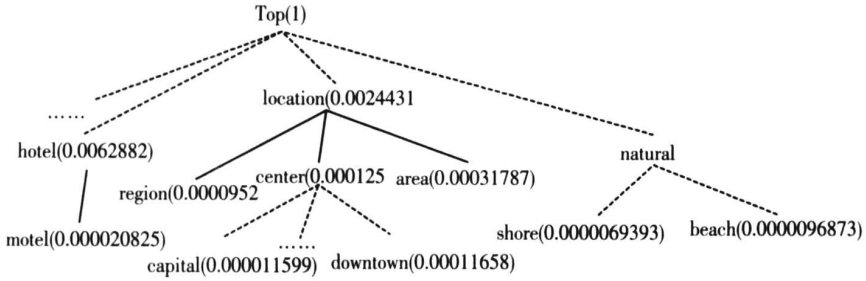

图 1　加权 IS-A 层次关系

Fig.1　IS-A level relation with weight

定义 **1**　加权 IS－A层次关系: 给定一个英文词汇库 ε, 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 $p(Top)=1$.

从上述加权 IS－A层次关系图中, 根据信息量的计算公式 $I(E)=-\log P_i$ 得到各个词汇的信息量, 比如: $I(\text{downtown})=-\log 0.000\,011\,658=4.933\,376$, $I(\text{area})=-\log 0.003\,178\,7=2.497\,751$. 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. 对于 2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 [7] 中提出的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 $ics(n_1, n_2)=\frac{2I(n')}{I(n_1)+I(n_2)}$; 其中, n' 为 n_1、n_2 的最大公共父节点词汇, 即 $I(n')=\max_{n\in S(n_1, n_2)}\{I(n)\}$, $S(n_1, n_2)$ 是所有 n_1、n_2 的公共父节点词汇.

1.2　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.

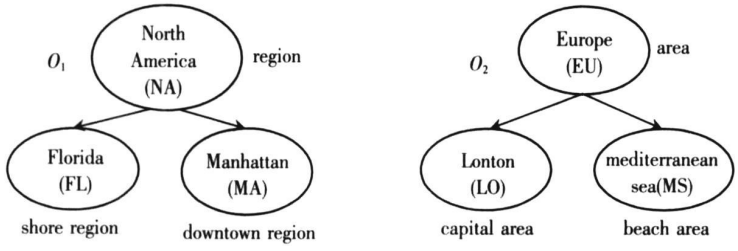

图 2　本体 O_1 和本体 O_2

Fig.2　Ontology O_1 and ontology O_2

提取本体 O_1 和 O_2 的属性, 根据词汇间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

Table 1　Semantic similarity matrix of attribute

ics	area	capital	beach
region	1	0.583 397	0
shore	0	0	1
downtown	0.703 002	0.790 786	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 (作为形式背景中的对象), 并且获得它们的属性 (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, 具体是根据以下思想来完善: 根据属性语义相似表得到每个属性间的 ics, 然后对于本体 O_1 中的每个概念, 在得到该概念拥有的所有属性在相似矩阵中组成的矩阵后, 取每一列的最大值并把它赋给该概念在本体 O_2 中对应属性的值, 用同样的方法处理本体 O_2 中的每个概念. 根据此思想获得的初步形式背景如表 2

表 2　初步形式背景

Table 2　Initial formal context

	region	shore	downtown	area	capital	beach
NA	1			1	0.583 397	0
FL	1	1		1	0.583 397	1
MA	1		1	1	0.790 786	0
EU	1	0	0.703 002	1		
LO	1	0	0.790 786	1	1	
MS	1	1	0.703 002	1		1

对于已经构造完成的形式背景, 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, 得到最终二值化形式背景, 利用 Godin算法[6]可以构建完整的概念格如图 3

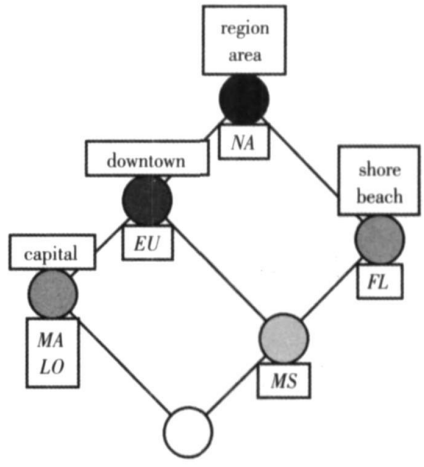

图 3　生成的概念格

Fig.3　Induced concept lattice

1.3　概念相似度量

本文采用文献[7]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进行衡量, 即只考虑形式概念的内涵.

定义 **3**　两形式概念 $C_1(A_1, B_1)$ 和 $C_2(A_2, B_2)$ 的相似度 $\mathrm{Sim}(C_1, C_2) = \dfrac{M(B_1, B_2)}{\max\{|B_1|, |B_2|\}}$, 其中, $M(B_1, B_2) = \max\left\{\sum_{b_1 \in B_1,\, b_2 \in B_2} ics(b_1, b_2)\right\}$, b_1 和 b_2 只能在每 1种组合中出现 1次.

1.4　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] 只提取了相等关系 (Equal) 和 父子关系 (Sub) 两种关系. 考虑概念MS和MA 2个节点, 他们具有除根节点以外的共同父节点, 这说明MA和MS存在相同的属性, 即是存在某种新的关系.

定义 **4**　如果概念格中 2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 $\mathrm{Overlap}(C_1, C_2)$. 其间相似度为 $\mathrm{Sim}(C_1, C_2)$.

根据定义 4, 发现存在交迭关系的有: $\mathrm{Overlap}(MS, MA)$, $\mathrm{Overlap}(MS, LO)$.

定义 **5**　如果概念格中 2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 $\mathrm{Sim}(C_1, C_2)$ 大于领域专家设定的一个阈值 λ, 那么这 2个概念之间存在的关系称之为相似关系, 记为 $\mathrm{Similarity}(C_1, C_2)$.

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

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.

2　结语

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.

[参考文献](References)

[1] Pan L Y, Song H, Ma F Y. A macrocommittees method of combining multistrategy classifiers for heterogeneous ontology matching[C]// Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2004, 3129: 672-677.

[2] Stoilos G, Stamou G, Kollias S. A string metric for ontology alignment[C]// Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2005, 3729: 624-637.

[3] Euzenat J, Valtchev P. Similarity-based ontology alignment in owl-lite[C]// Proc of the European Conference on Artificial Intelligence. Valencia: IOS Press, 2004: 333-337.

[4] Castano S, Ferrara A, Montanelli S. Matching ontologies in open networked systems: techniques and applications[J]. Journal on Data Semantics V, Lecture Notes in Computer Science, 2006, 3870: 25-63.

[5] Liya Fan, Tianyuan Xiao. FCA-mapping: a method for ontology mapping[C]// Proc of the 4th European Semantic Web Conference. Berlin, Heidelberg: Springer-Verlag, 2007: 192-206.

[6] Godin R, Missaoui R, Alaoui H. Incremental concept formation algorithms based on galois (concept) lattices[J]. Computational Intelligence, 1995, 11(2): 246-267.

[7] Anna Formica. Concept similarity in formal concept analysis: an information content approach[J]. Knowledge-Based Systems Archive, 2008, 21(1): 80-87.

[责任编辑: 顾晓天]

(上接第 76页)

[8] 曲维光, 吉根林, 穗志方, 等. 基于语境信息的组合型分词歧义消解方法[J]. 计算机工程, 2006, 32(17): 74-76.
Xiao Yun, Sun Maosong, Benjamin K Tsou. Solving combinatorial ambiguity in Chinese word segmentation using contextual information[J]. Computer Engineering and Application, 2001, 37(19): 87-81. (in Chinese)

[9] 冯素琴, 陈惠明. 一种自组织的汉语组合型歧义消歧方法[J]. 计算机工程与设计, 2007, 28(3): 737-749, 742.
Feng Suqin, Chen Huiming. AdaptiveChinese combinatorial ambiguities disambiguate method[J]. Computer Engineering and Design, 2007, 28(3): 737-749, 742. (in Chinese)

[10] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 18th ICML. San Francisco: Mogan Koufmann, 2001: 282-289.

[11] 冯素琴, 陈惠明. 基于语境信息的汉语组合型歧义消歧方法[J]. 中文信息学报, 2007, 21(6): 13-16, 42.
Feng Suqin, Chen Huiming. Context-based approach to combinational ambiguity resolution in Chinese word segmentation[J]. Journal of Chinese Information Processing, 2007, 21(6): 13-16, 42. (in Chinese)

[责任编辑: 孙德泉]