本体映射框架的设计*)

郑丽萍[1] 李光耀[1] 梁永全[2]

(同济大学电子信息工程学院 上海 201804)[1] (山东科技大学信息学院 青岛 266510)[2]

摘 要 本体异构是本体间互操作的主要障碍。解决本体异构最好的方法是本体映射。本文主要讨论了本体异构和本体映射的研究现状,提出一种具有混合体系结构的本体映射框架,并用具体的本体映射来说明该映射框架。

关键词 本体,本体映射,WordNet

Design of Ontology Mapping Framework

ZHENG Li-Ping[1] LI Guang-Yao[1] LIANG Yong-Quan[2]

(College of Electron and Information Engineering, Tongji University, Shanghai 201804)[1]

(College of Information Science and Engineering, Sdust, Qingdao 266510)[2]

Abstract Ontology heterogeneity is the primary obstacle for interoperation of ontologies. Ontology mapping is the best way to solve this problem. In this paper, ontology heterogeneity and the current study situations of ontology mapping are discussed, and an ontology-mapping framework with a kind of hybrid architecture is put forward. An example of ontology mapping is given to explain the mapping framework.

Keywords Ontology, Ontology mapping, Wordnet

1 引言

本体(Ontology)一词来源于哲学领域,主要研究物质的性质及其内在关系。目前,本体应用在智能信息集成、协作信息系统、信息检索、电子商务和知识管理等领域。本体的应用领域越来越多,但应用的主要目的都是为了知识共享和复用。

许多学科都使用本体这个概念,但却存在不完全相同的定义和理解。本体的定义有许多种,定义之间的侧重点也各不相同,但本体的本质是对共享概念的一个正规清晰的描述。由于本体自身的分散特性,并且本体的构造还没有一个统一的标准,不同的用户可以构造不同的本体,因此导致了在同一个或者重叠的领域产生了许多不同的本体,即使一个小的背景领域也可能出现许多不同的本体。在同一领域内,要想实现不同本体间的互操作就必须解决本体间的异构问题。

2 基本概念

2.1 本体(ontology)

本体最初是一个哲学概念,用来描述事物的本质[1]。在20世纪80年代,科研人员把本体引入人工智能领域,并赋予其新的含义。在计算机科学领域,本体被定义为共享概念模型的形式化规范说明[3]。由于本体的分类方法很多,目前还没有能够被广泛接受的分类标准。一般认为本体包含5个基本建模原语:类、关系、函数、公理和实例。科研人员从实际出发提出多种构造本体的标准,其中最有影响的是T. R. Gruber提出的5个准则[2]:明确性和客观性、一致性、完全性、最大单调可扩展性、最小承诺。目前广泛被使用的本体主要有Wordnet、Framenet、GUM、SENSUS、MIKROKMOS[4]。

2.2 本体映射(ontology mapping)

本体映射确定不同的本体怎样被映射或者怎样被相互关联。它是本体间概念和关系取得一致性的一个规范说明。本体映射与模式匹配有相似之处。本体的映射一般要经过五步:信息本体化、相似性的提取、语义映射、映射执行和映射后处理[5]。本体映射的关键是概念间相似度的计算。首先定义相似度:Sim:$w_1 \times w_2 \times o_1 \times o_2 \rightarrow [0,1]$,相似值在0和1之间。Sim(A,B)表示A和B之间的相似度。w_1 和 w_2 是两个本体所基于的术语集,O_1 和 O_2 是需要映射的两个本体。

——Sim$(e,f)=1$:表示概念 e 和概念 f 是相同的两个概念;

——Sim$(e,f)=0$:表示概念 e 和概念 f 是两个完全不同的概念。

2.3 Wordnet简介

Wordnet是一个在线字典参考系统。它是基于心理语言规则而设计的英文词典,并以synsets为单位组织信息。所谓synsets是把在特定的上下文环境中可以互换的英语名词、动词、形容词和副词组织成一个同义词集合。Wordnet的每一个词集都表示一个潜在的单词概念。每个同义词集不仅包含一系列同义词,还包含同义词的扩展以及该同义词集与其他同义词集关系的描述。

3 解决本体异构的方法

为了实现异构本体间的互操作,一般可采用三种方法[6]:它们的体系结构如图1所示:

(1)本体间建立包含关系。目标本体简单地包含源本体,来自源本体的所有数据概念都能在目标本体中出现。该方法的缺点是信息和概念只能被复用而不能被修改。

(2)本体间建立映射关系。本体映射就是概念层上语义相关的实体根据语义关系进行转换的过程。通过映射源本体的实体可以转换成目标本体的实体。

(3)建一个公共的本体。把多个源数据所对应的本体进行合并,生成一个完整的公共本体。也就是寻找一个在任何

*)基金项目:国家自然科学基金资助项目(70371052)。郑丽萍 博士生,主要研究领域:图形图像;李光耀 教授,博导;梁永全 教授,博导。

情况下用户都能进行查询的全局本体。全局本体为具体的语义说明提供了一个共享的词汇表。所有系统或信息资源所对应的本体都连接到全局本体上，因而它们的语义是一致的。这种方法的难度较大，不易实现。

因此解决本体异构最有效的方法就是本体间进行映射。本体映射的目的就是找到本体中概念之间的对应关系，并制定出相应的映射规则。

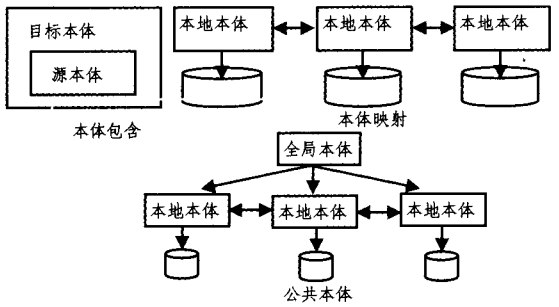

图 1　处理本体异构问题的方法

4　本体映射的研究现状

目前国外许多著名的大学和实验室都对本体映射进行了研究，并且一些具体的映射系统和实现方法已经被开发出来，例如 Glue 系统[7]、本体代数方法[8]、MAFRA[9]、IF-Map[10] 和 OMEN[11] 系统。虽然科研人员一直在研究本体映射并解决了一些冲突，但是还存在一些不足。

目前，计算两个本体 O_1 和 O_2 中概念的相似度时，本体中的每一对概念都被考虑在内。如果本体 O_1 中含有 m 个概念，本体 O_2 中含有 n 个概念，那么就要计算 $m\times n$ 次相似度，也就是每对概念之间的相似度都要计算出来，并形成 $m\times n$ 维的相似矩阵，因此计算量很大。有的两个概念根本就不相似，所以计算它们的相似度是不必要的，只会增加时间复杂度和空间复杂度。因此计算时应该对概念对的数量进行限制，以减少相似度的计算量。为了减少概念对的数量，本文提出一个改进的本体映射框架。

5　本体映射框架 MOMF

本体的映射是在多个本体间进行映射。在多本体的环境中，每一个信息源都有各自的本地本体，但它们使用的术语表可能是不同的，这些本体间的关系是松散耦合的。本文在多个本体环境中，加入一个共享领域词汇表，形成一种混合的体系结构。在混合体系结构中，每个信息源都有自己的本地本体，但本地本体的建立都以一个共享领域词汇表为参考。

根据上述设计思路，设计了一个改进的本体映射框架——多方法本体映射框架 MOMF(Multiple-way Ontology Mapping Framework)。该映射框架由 5 部分组成：

(1)一个应用本体(application ontology)：应用本体包含一个已经存在的上层(top-level)本体，本文使用 Wordnet 系统。它提供概念术语的一些同义词和近义词。

(2)参考层(reference layer)：它包括概念术语的数据类型匹配表和相关信息说明。数据类型匹配表定义了各个类型之间的匹配程度。不同数据类型间的匹配度在 0 和 1 之间。相关信息说明包括一些单词的缩写、常见的缩写词等，如 NO，NUM。

(3)共享领域词汇表(shared domain vocabulary)：它是一个共享领域的全局术语词汇表。共享领域词汇表根据 Wordnet 来分类。表的每一行都是某一概念术语的近义词或者同义词。计算概念相似度时，以该词汇表为参考只计算同义词和近义词之间的相似度。先根据 Wordnet 系统设置同义词和近义词的初始相似度，其它概念对的相似度赋值为 0，并初始化概念的相似矩阵。

(4)各个本地本体：不同的用户根据自己的需要而建立的本地本体，每个本体的建立都参考共享领域词汇表，因此这些本体中的概念在语义上就存在一定的关系。

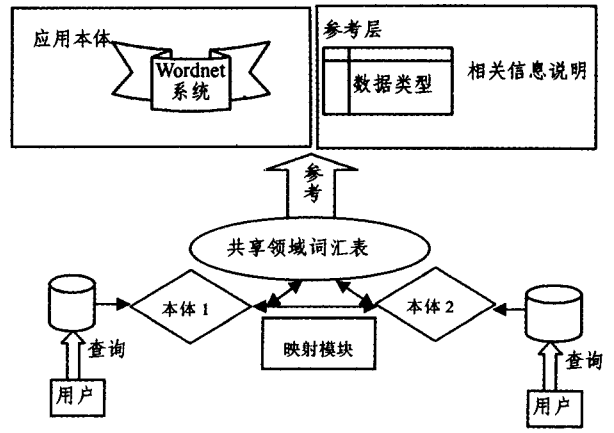

图 2　MOMF 结构

(5)映射模块：该模块的任务是进行概念相似度的计算。在计算时，使用改进的相似度计算方法分别计算语义相似度和描述相似度。改进的相似度计算方法是把基于实例的方法和启发规则方法相结合。语义相似度是指概念之间自身语义的相似程度；描述相似度就是从属性和关系的角度说明概念的相似性，是指概念的属性或概念间关系的相似程度。该映射框架的结构如图 2 所示。

6　具体实例应用

比如本体 O_1 和本体 O_2 都是关于学校课程和老师情况的本体，它们都以共享领域词汇表为基础建立。但本体 O_1 和本体 O_2 是异构的。本体 O_1 和本体 O_2 的结构如图 3 所示：

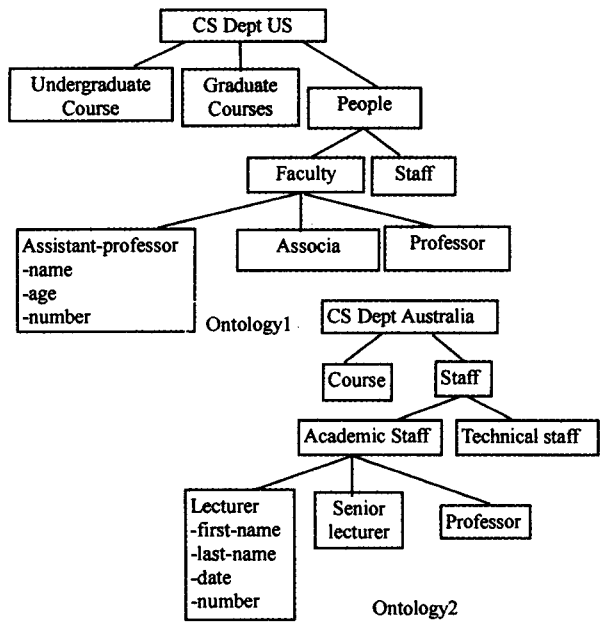

图 3　本体结构树

(下转第 180 页)

每条序列运行3次得到的最好结果。IMC算法所示数据为表2利用IMC每条序列运行3次得到的最好结果。MC算法所示数据为文[17]报告的最小能量。通过利用新运动集和传统运动集在MC算法得到的结果比较可知，IMC方法在10条序列中有6条能找到较MC算法更低的能量构象，有2条串能得到相同的能量构象，只有两条序列所得结果要差于MC算法。

结论 利用方形网格模型，我们应用一种扩展的运动集，并与传统的运动集进行了比较，实验结果表明改进的遗传算法IGA能在最长的两条二维HP序列上比以前的算法找到更好的解。相比较而言，IGA方法在二维序列的7号串和8号串上均找到了目前已知的最优解，而前面的所有进化算法只是找到了次最优解。改进的遗传算法是目前格子模型中优化算法较为有效的一种算法。可以得出这样的结论，蛋白质折叠的优化过程与所选用的运动集有很大的关系，在折叠过程中，使用更为灵活的运动集能更快更好地找到最低能量构象。因此可以这么说，改进的遗传算法尤其适用于再现蛋白质折叠的折叠途径，同时我们也希望该运动集在处理蛋白质折叠的模拟过程中起到一定的作用。

参考文献

1 Berger B, Leighton T. Protein folding in the hydrophobic-hydrophilic model is NP complete, J. Comput. Biol, 1998(5):27～40
2 Paterson M, Przytycka T. On the complexity of string folding, Discrete. Appl. Math., 1996,71: 217～230
3 Kirkpatrick S, Gelatt Jr C D, Vecchi M P. Optimization by simulated annealing. Science, 1983,220:671
4 Unger R, Moult J. Genetic algorithms for protein folding simulations. J Mol Biol, 1993, 231:75
5 Konig R, Dandekar T. Improving genetic algorithms for protein folding simulations by systematic crossover. Biosystems, 1999, 50: 17～25
6 Liang Faming. Evolutionary Monte Carlo for protein folding simulations. J Chem Phys,2001, 115:3374
7 Jiang Tianzi. Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. J Chem Phys, 2003,119:4592
8 Beutler T C, Dill K A. A fast conformational search strategy for finding low energy structures of model proteins. Protein Science, 1996,5:2037～2043
9 Zhang J L, Liu J S. A new sequential importance sampling method and its application to the two-dimensional Hydrophobic- Hydrophilic model. J Chem Phys, 2002, 117(7): 3492～3498
10 Grassberger P. The pruned-enriched Rosenbluth method: simulations of Theta polymers of chain length up to 1000000. Phys Rev E, 1997, 56(3): 3682～3693
11 Shin J, Oh W S. Study of move set in cubic lattice model for protein folding. J Phys Chem, 1998, 102(33): 6405～6412
12 Nunes N L, Chen K, Hutchinson J S. A flexible lattice model to study protein folding. J Phys Chem, 1996, 100(24): 10443～10449
13 Yesylevskyy S O, Demchenko A P. Towards realistic description of collective motions in the lattice protein folding models. Biophysical Chemistry, 2004, 109(1): 17～40
14 Konig R, Dandekar T. Improving Genetic Algorithms for Protein Folding simulations by systematic crossover. Biosystems, 1999, 50(1): 17～25
15 Li H, Tang C, Wingreen N. Nature of driving force for protein folding: A result from analyzing the statistical potential. Physical review letters,1997, 79(4): 765～768
16 Błażewicz J, Łukasiak P. Application of tabu search strategy for finding low energy structure of protein. Artificial Intelligence in Medicine, 2005, 35(1-2):135～145
17 Yue K, Fiebig KM. A test of lattice protein folding algorithms. Proc Natl Acad Sci USA,1995, 92(1): 325～329

(上接第167页)

如果本体O_1的用户要在本体O_2中查找与本体O_1中的某个概念相关的信息，那么需要在O_1和O_2之间进行映射。以Wordnet为参考，得到概念的一个共享近义词表，如表1所示。利用共享近义词汇表，获得近义词间的最初相似度，然后再进一步计算相似度。

表1 共享近义词表

course	undergraduate courses, graduate courses, postgraduate courses,
people	staff, faculty, academic staff, technical staff, teacher
name	frist -name, last-name
job	professor, assistant-professor, associated-professor, lecturer, senior lecturer,

表2 数据类型匹配

匹配值	实型	整型	字符型	日期型
实型	1	0.9	0.1	0.7
整型	0.9	1	0.1	0.8
字符型	0.1	0.1	1	0.1
日期型	0.7	0.8	0.1	1

在参考层中，设定一个数据类型匹配表，如表2所示。相似度的计算由映射模块来完成。最后，输出相似矩阵。设定一个阈值，当相似度大于该值则认为两个概念相似并更新相似矩阵，否则这两个概念不相似并且相似矩阵保持不变。

总结与展望 两个本体间的映射是1：1，而多个本体间的映射有1：n和m：n。对于m：n关系的映射，可以分解成m个1：n关系的映射。让这m个用户按一定的优先级排队等待并进行查询。每个用户进行查询时就可以按1：n关系进行映射和查询。m个用户的查询是串行进行的。排队的顺序可以是先进先出或其它的排队方法。多个本体间的1：n映射可以转换成n对两个本体间的映射。这n对两个本体间的映射可以同时进行。

随着计算机的发展，本体的应用领域越来越多，本体的数量也越来越多。总的来说，本体的研究和应用还处于起步阶段，许多问题还需要近一步的研究。另外，本体还没有统一的生命周期定义和标准，也没有统一的本体开发的方法学和技术。因此，创建系统的、全面的、完整的方法体系仍是本体未来的研究方向。

参考文献

1 Kivela A, Hyvonen E. Ontological theories for the Semantic Web [M], Helsinki: HIIT Publications,2002. 111～136
2 Gruber T. Towards principles for the design of ontologies used for knowledge sharing [J]. International Journal of Human-Computer Studies,1995,43(5-6): 907～928
3 Gruber T R. A translation approach to portable ontology specification [J]. Knowledge Acquisition, 1993, 5(2):199～220
4 邓志鸿，唐世渭，张铭，等. Ontology研究综述[J]. 北京大学学报(自然科学版)，2002,38(5):730～738
5 Maedche A, Motik B. Ontologies for Enterprise Knowledge Management [J]. IEEE Intelligent Systems,2003. 26～33
6 Ehrig M, Sure Y. Ontology Mapping - An Integrated Approach [J]. In: Proceedings of the 1st European Semantic Web Symposium, Heraklion,Greece, Springer, LNCS,2004. 10～12
7 Doan A, Madhavan J, Domingos P. Learning to Map between Ontologies on the Semantic Web [J]. In: Proc. World-Wide Web Conf. ACM Press, May 2002. 662～673
8 Wiederhold G. An algebra for ontology composition [D], U. S. Naval Postgraduate School, Monterey CA, 1994
9 Macedche A, Motik B. MAFRA——A Mapping Framework for Distributed Ontologies [J]. Web Intelligence and Agent System, 2003,1: 235～248
10 KalfoglouY, SchorlemmerM. Information-flow-based ontology mapping [J]. In: Proceedings of the 1st International Conference,Springer, 2002. 1132～1151
11 Mitra P, Noy N F, Jaiswal A R. OMEN: A Probabilistic Ontology Mapping Tool [J]. In: Workshop on Meaning coordination and negotiation at the Third International Conference on the Semantic Web (ISWC-2004), Hisroshima, Japan