语义集成:本体映射方法研究

李选如,何洁月
(东南大学 计算机科学与工程学院,江苏 南京 210096)

摘 要:本体是客观世界知识的表现形式,随着语义 Web 研究的深入,研究者们构建了越来越多的本体,如何实现本体之间的知识共享和重用,成为了语义 Web 发展的关键。文中对本体映射的方法进行了研究,系统阐述了本体及本体映射的定义、本体映射中的相似度计算和本体映射框架等。如何减少本体映射中的人工干预,实现本体的半自动化或自动化映射将是该领域的发展方向。

关键词:本体;本体映射;语义 Web
中图分类号:TP301.2 **文献标识码**:A **文章编号**:1673-629X(2007)02-0121-04

Semantic Integration: Survey of Ontology Mapping Approaches

LI Xuan-ru, HE Jie-yue
(Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract: Ontology is an explicit specification of a conceptualization. With the development of semantic Web, there are more and more ontologies. How to achieve the reusing and sharing of knowledge between ontologies is a key in the development of semantic Web. In this paper, we give a survey of the existing approaches about ontology mapping, including the definition of ontology and ontology mapping, the approaches of similarity calculation, and some existing frameworks of ontology mapping. How to reduce the labor interference in ontology mapping is a working direction in the future.

Key words: ontology; ontology mapping; semantic Web

0 引 言

近年来,随着语义 Web 的发展,本体的研究也在不断增强,研究者们创建了越来越多的本体。如何协调这些本体之间的知识共享和重用问题,使其能够更好地应用于语义 Web,已经迫在眉睫。本体映射这个概念随之产生,初始阶段,研究者们只能局限于手工完成本体之间的映射,然而,手工构建本体之间的映射是一个十分繁杂的过程,因此,如何半自动或自动构建本体之间的映射就显得尤为重要。本体映射的自动化构建涉及到的领域非常广,其中包括机器学习、自然语言处理、人工智能等,因此,本体映射的半自动或自动构建也是当今研究的一个难点。

1 本体和本体映射

本体最早是一个哲学上的概念,从哲学的范畴来说,本体是客观存在的一个系统的解释或说明,关心的是客观显示的抽象本质。Gruber[1] 给出了本体的最为流行的定义,即‘本体是概念模型的明确的规范说明”,其形式化定义为:$OT = < \mathrm{Id}, C, R, F, T, I >$,其中 Id 是本体标识;$C$ 为概念集合;R 是领域中概念之间的关系集,$R: C_1 \times C_2 \times \cdots \times C_n$,基本的关系是子类关系(subclass-of),部分整体关系(part-of),继承关系(kind of)以及属性关系(attribute-of);F 是函数关系,该关系的前 $n-1$ 个元素可以唯一决定第 n 个元素,$F: C_1 \times C_2 \times \cdots \times C_{n-1} \rightarrow C_n$;$T$ 代表永真断言;I 是实例,从语义上讲实例表示的就是对象。

本体映射是指在两个本体中存在语义级的概念关联,通过语义的联系,实现将源本体的实体(概念、实例、属性等)映射到目标本体实体上的过程。文献[2]提到:本体映射就是指给定两个本体 A 和 B,对于 A 上的每一个实体,设法在 B 上找到与其有相同或相近语义的实体,这些实体包括本体中的类、属性以及类的实例。Ehrig[3] 给出了一个形式化的本体映射函数:

map: $O_1 \rightarrow O_2$;

如果 $\mathrm{sim}(e_1, e_2) > s$, $\mathrm{map}(e_1) \rightarrow e_2$,其中,$e_1$ 和 e_2

收稿日期:2006-05-18
基金项目:江苏省高技术研究计划(G2004034)
作者简介:李选如(1980-),女,江苏镇江人,硕士研究生,研究方向为数据库技术、数据挖掘、信息集成;何洁月,副教授,研究方向为数据库技术、数据挖掘、信息集成。

分别是两个本体中的实体，$\mathrm{sim}(e_1,e_2)$ 是这两个实体之间的相似度，s 是相似度阈值。

Ehrig and Staab[4] 总结了过去的工作，归纳出本体映射的 6 个过程：

(1)特征提取：提取用于计算相似度的特征，如概念、属性的名称等；

(2)选择用于映射的概念对；

(3)进行相似度计算；

(4)相似度整合：通常有多种方法可以衡量本体实体之间的相似度，得出多种相似度值，因此要对各相似度进行综合考虑，从而得到一个整体上的相似度；

(5)优化：第(4)步结束后，已经得到待映射的各个实体之间的初始相似度，这时一般需要人工的干预，利用领域知识，对其进行调节；

(6)迭代第(1)步到第(5)步，直到达到满意结果。

2 本体映射相似度计算方法

通过本体的映射过程，可以看出，映射最关键的就是计算本体实体之间的相似度，从而发现它们之间的语义联系，最终完成映射。本体实体之间的相似度充分体现了客观世界中各种事物之间的联系，如何计算这些相似度，笔者认为主要可以从两个方面来考虑相似度的计算：一方面是从元素层来考虑各实体之间的相似度；另一方面是从本体结构上的特征来考虑。

2.1 基于元素层的相似度

基于元素层的相似度，是指在进行本体映射时，只考虑本体中各元素(如概念、实例等)之间的相似程度，而不用从整体结构上考虑(例如概念之间的关系等)。具体而言，就元素层的相似度计算方法又可分为如下 3 类：

(1)基于字符串匹配。

基于字符串匹配主要考虑本体中各元素名称的相似程度，基于“两个字符串越相似，则这两个概念越类似”的思想，通过已有的字符串匹配算法，得到相似概念集。该技术主要考虑字符串的下列特征：前缀、后缀以及编辑距离[5,6]等。

(2)基于元素的约束。

基于元素的约束主要考虑本体中各元素定义时的内部约束特征，如数据类型、属性个数等。例如，本体 A 中定义了一个概念“MONTH”，它的数据类型是“datatype”，另一个本体 B 中也定义了两个概念“MONTH”，但是一个类型是“datatype”，另一个是“int”，则可以说 A 中的“MONTH”匹配 B 中类型为“datatype”的“MONTH”，而不是后者。

(3)基于语义的匹配技术。

语义问题是本体映射中特有的问题，就本体中单个概念而言，出现语义问题，主要是指同词异义，或异词同义。针对这种情况，可以考虑利用现有的一些词典来处理。例如，文献[7]采用了一种从 WordNet 中获得同义词的方法，用于解决上述问题。

2.2 基于结构层的相似度

基于结构层的映射技术，主要从实体之间的相互关系上考虑。在本体映射时，从本体的整体结构考虑本体中实体之间的相似程度。

(1)基于图论的方法。

该方法中，待映射的本体以图或树的形式表示，其中，一个本体是一张图(树)，本体中的每个实体(类、属性、实例等)表示为图(树)中的结点，实体之间的关系(subClassOf，superClassOf 等)以图(树)中结点之间的连线来表示。通常，考虑两个本体中实体的相似性，以各自在图中所处的位置来决定。

下面就此方法的几个关键技术作一说明：

a. 图形匹配：通过本体中的实体在图中的具体位置、相互之间的路径等因素来考虑其相似性，典型的应用可参考文献[3]。

b. 子结点：在衡量两个实体结点相似性时，考虑它们各自的子结点是否相似，如果相似，则这两个概念也相似。更详细的描述可参考文献[5]。

c. 叶子结点：通过直接比较非叶子结点所拥有的叶子结点之间的相似性，来衡量它们的相似性。该方法与前者的区别是，不用考虑结点的直接子结点，具体的实现可参考文献[5,8]。

(2)基于分类学的方法。

该方法与上面提到的基于图论的方法类似，仅考虑待映射本体的实体之间的联系。通过定义若干的规则，得出本体之间结构上的相似度。例如，如果两个类的父类相似则得出这两个类之间也存在相似性；同样也可以得出它们的兄弟类之间也存在一定的相似性[9,10]。

3 本体映射系统

上一小节主要对本体映射过程中的相似度计算作了总结，如何将这些相似度运用到本体映射中，是本体映射框架研究的重要方面。目前的本体映射框架都在一定程度上实现了本体映射的半自动化构建。

3.1 GLUE

GLUE[1] 在进行相似度计算时，不是针对其中特定的相似度进行比较，而是通过计算概念之间的联合分布概率来评估它们之间的相似度。该系统首先需要计算 4 组概率分布值，即 $P(A,B)$、$P(A,\bar{B})$、$P(\bar{A},$

$B)$、$P(\bar{A},\bar{B})$,其中 $P(A,B)$ 定义为一个实例同时属于概念 A 和 B 的概率,$P(A,\bar{B})$ 定义为一个实例属于概念 A 而不属于概念 B 的概率,其余的依次类推;然后通过相似度函数

$$\text{Jaccard}-\text{sim}(A,B)=P(A\cap B)/P(A\cup B)=\frac{P(A,B)}{P(A,B)+P(A,\bar{B})+P(\bar{A},B)}$$

计算各概念的相似度。

该系统主要由三层结构构成:

(1)分布估计层(Distributer Estimator),由一个元学习器和一组基本学习器组成,通过多策略学习方法,计算出四组概率值。

(2)相似度估计层(Similarity Estimator),利用相似度函数,对上层输出的概率值,计算各概念对的相似度。

(3)松弛标签(Relaxation Labeler),主要是从本体结构上考虑,参考相关领域知识,对结果进行优化。

3.2 Prompt 和 Anchor Prompt

Prompt[11]是一个交互性的本体合并工具。首先,利用字符串匹配技术给出一个初始的匹配本体集,供用户选择;然后根据用户的选择,执行本体的匹配,途中如果发现冲突,会给出一些建议,供用户再次选择,如此循环,直至最终生成映射好的本体。

Anchor Prompt[11]是在 Prompt 的基础上提出的,它将每个本体看成一个图,本体中的每一个实体看成图中的一个结点,实体之间的关系,则在图中表示为结点之间的联系。该方法通过分析图中结点出现的位置,以及结点之间的路径,来确定两个本体中的某些实体是否相似。

3.3 MAFRA

MAFRA[12]将待映射的本体统一为 PDF(S)格式,采用多策略的学习方法,并参考 WordNet 的同义词集,来计算本体中概念的相似度。MAFRA 构建了一个语义桥本体,该本体封装了一切用于源本体和目标本体进行映射的信息。通过这个本体,来完成源本体和目标本体之间的映射。另外,MAFRA 也借鉴了 Prompt 中的方法,定义了一个人机交互的界面,用户可以干预映射过程。

3.4 FCA-Merge

FCA-Merge[13]使用基于形式概念分析(Formal Concept Analysis)的技术,产生一个与本体中概念相关的概念栅格,并对其分析,完成本体之间的映射。该系统主要由三部分构成:

(1)文本的语义分析,该层的输入是两个待合并的本体,以及一组与这两个本体都相关的文档,使用自然语言处理技术,从这组文档中提取与各本体相关的概念,为每个本体形成一个二维表格,反映每个本体中的概念在某个文档中是否出现。

(2)概念栅格的形成:对上一步的输出进行分析,如果两个本体中的某些概念总是出现在同一些文档中,则这两个概念相似,进行概念合并。

(3)本体构建:根据上一层的输出,构建本体。由于本体的构建需要丰富的领域知识,该层需要人工的干预。

3.5 BayesOWL

BayesOWL[14]提供了一组规则和方法,可直接将一个用 OWL 表示的本体转换为 BN 结构。此外,还提供了一个基于 IPFPD 方法,利用类之间的关联,约束构建 CPT 表。Rong Pan[15]利用 BayesOWL 建立了一个可以自动完成本体映射的模块。该模块由三部分组成:

(1)用于文本分类的学习器,直接从互联网上获取与输入本体中概念相关的文档,进行学习,从而得到一组概率分布;

(2)BayesOWL 模块:用于将给定的本体转换为 BNs 表示的结构;

(3)概念映射模块:基于 BNs 完成本体之间的映射。

3.6 NOM 与 QOM

NOM[3]分别从元素层与结构层上考虑,定义了 17 条规则,计算各种相似度,然后对各种相似度进行加权求和,得到最后的相似度,并据此完成本体之间的映射。QOM[4]是在 NOM 基础上构建的,并对 NOM 的性能进行了优化,例如,在计算相似度时,NOM 中是对本体中每个实体对都进行测试,而 QOM 中,通过特定算法,有选择地进行实体对的相似度计算。

3.7 GMO

针对已有的本体映射方法大都是基于本体概念之间的语义相似性(主要通过概念实体的 Label 值)来完成初始的映射,并在此基础上进行结构相似度的计算,Wei Hu[16]提出了一种直接从结构上来考虑本体之间映射的方法,即 GMO 模型。GMO 中首先将本体转换成双边图(Bipartite Graph),通过评价图形向量之间的相似度,完成本体之间相似度的计算。

4 结论与展望

本体是人和机器,程序间知识交流的基础,构建本体的根本目的就是为了知识的共享和重用。如何减少本体之间存在的语义冲突,以更好地实现本体映射,已经成为本体研究领域的重要课题。目前的本体映射框

架都在一定程度上实现了本体映射的半自动化构建。Prompt 和 Anchor Prompt 是一个人机交互式的本体映射工具,其结果很大程度上依赖于人的决策。GLUE 将机器学习的方法应用于本体映射,通过对本体自身实例的学习来完成本体的构建,减少了人工的干预,但对于实例不多的本体,其学习的效率将受到很大的制约。BayesOWL 和 FCA - Merge 直接从互联网上获得与待映射本体领域相关的资源,运用自然语言处理技术、机器学习等方法,完成本体映射,因而如何更好地获得与待映射本体相关领域的资源是决定其效果好坏的关键。NOM 和 QOM 则直接利用本体自身的特点,通过结合元素层相似度和结构层相似度完成映射,因此使用这两方法进行映射的本体必须是结构完好的。基于目前大多数本体映射框架都是通过先对本体中的实体的词义的分析,获得本体实体的元素层的相似度,然后基于元素层的相似性,得出结构层的相似性,GMO 提出了将本体转换成双边图,直接通过图形的匹配,即直接从本体的结构上完成本体实体的相似度的计算,实现本体映射,与 QOM 和 NOM 类似,使用 GMO 方法映射的本体,其结构应该是完好的。

综上所述,本体映射是个相当艰巨的工程,不仅要利用本体自身所反映的语义信息,还需要利用词典、互联网等外部资源;由于本体是由自然语言描述的,因而本体映射必然要使用到自然语言处理技术、机器学习、概率统计等,完整的本体映射框架如图 1 所示。如何充分利用各种外部资源、如何构建结构完好的本体以及如何更好地将自然语言处理技术、机器学习、概率统计等运用到本体映射上,从而减少人工干预,将是未来本体映射的发展方向。

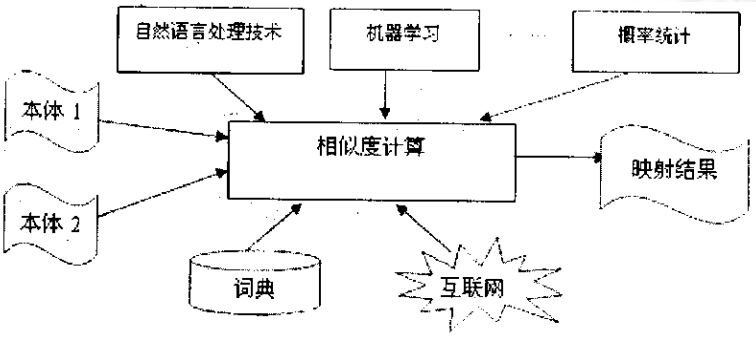

图 1 本体映射框架示意图

参考文献:

[1] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisiiion,1993,5:199 - 220.

[2] Su X. A text categorization perspective for ontology mapping [R]. Norway:Norwegian University of Science and Technology,2002.

[3] Ehrig M,Sure Y. Ontology mapping - an integrated approach [EB/OL]. 2005 - 02 - 14. http://www. aifb. uni - karlsruhe. de/WBS/ysu/publications/2004- mapping-TR. pdf/.

[4] Ehrig M,Staab S. QOM - Quick Ontology Mapping[C]//In: ISWC 2004,LNCS 3298. [s. l.]:[s. n.],2004:683 - 697.

[5] Do H H,Rahm E. COMA - A system for flexible combination of schema matching approaches[C]//In:Proceedings of the Very Large Data Bases Conference. Roma, Italy:[s. n.], 2001:610 - 621.

[6] Giunhiglia F, Shvaiko P, Yatskevich M. Semantic schema matching[R]. Trento:University of Trento,2005.

[7] Doan A,Madhavan J,Domingos P,et al. Learning to map between ontologies on the semantic web[C]//In:Proceedings of the 11th International Conference on World Wide Web. Hawaii,USA:[s. n.],2002:662 - 673.

[8] Madhavan J,Bernstein P A,Rahm E. Generic schema matching with cupid[C]//27th VLDB Conference. USA:Morgan Kaufmann Publishers,2001:49 - 58.

[9] Maedche A,Staab S. Measuring Similarity between Ontologies [C]//In:Proceedings of the European Conference on Knowledge Acquisition and Management EKAW - 2002. [s. l.]:[s. n.],2002:251 - 263.

[10] Melnik S,Garacia - Molina H,Rahm E. Similarity flooding:A versatile graph matching algorithm[C]//In Proceedings of the International Conference on Data Engineering (ICDE). San Jose,California:[s. n.],2002:117 - 128.

[11] Noy N F,Musen M A. The Prompt suite:Interactive tools for ontology merging and mapping[J]. International Journal of Human - Computer Studies,2003,59(6):983 - 1024.

[12] Maedche A, Motik B, Silva N, et al. MAFRA - a mapping framework for distributed ontologies[C]//In 13th European Conference on Knowledge Engineering and Knowledge Management EKAW. Madrid,Spain:[s. n.],2002.

[13] Stumme G,Maedche A. FCA - Merge:Bottom - up merging of ontologys[C] // In 7th Intl. conf. on Artificial Intelligence (IJCAI '01). Seattle,WA:[s. n.],2001:225 - 230.

[14] Ding Z,Ping Y,Pan Rong. A Bayeian Approach to Uncertainty Modeling in OWL Ontology [C]//In Proceedings of 2004 International Conference on Advances in Intelligent Systems - Theory and Applications (AISTA2004). Kircherg,Luxembourg:[s. n.],2004.

[15] Pan R,Ding Z,Yu Y,et al. A Bayesian Network Approach to Ontology Mapping[C]//Proceedings of the 4th International Semantic Web Conference. Ireland:Springer,2005:1 - 15.

[16] Hu W,Jian N,Qu Y,et al. GMO:A Graph Matching for Ontologies. Submitted to KCap workshop on Integrating Ontologies[C/OL]. 2005. http://xobject. seu. edu. cn/project/Falcon/GMO. pdf.