基于机器学习的本体集成框架*)

欧 灵[1,2] 张玉芳[1] 吴中福[1] 钟 将[1]

(重庆大学计算机学院 重庆 400044)[1] (西南大学计算机科学系 重庆 400715)[2]

摘 要 语义服务是下一代Web服务面临的关键问题,语义网为实现广泛的语义服务提供了可能,Ontology是语义网体系结构的核心。针对协作的分布式系统需要语义互联的问题,本文分析了造成语义互联困难的主要因素是本体的匹配和集成,提出了一个基于机器学习的Ontology集成的框架模型。

关键词 本体,语义网,集成

Framework of Ontology Integration Based on Machine Learning

OU Ling[1,2] ZHANG Yu-Fang[1] WU Zhong-Fu[1] ZHONG Jiang[1]

(College of Computer, Chongqing University, Chongqing 400030)[1]

(Dept. of Computer Science, Southwest University, Chongqing 400715)[2]

Abstract Semantic service is the key problem existing in Web service of next generation, and semantic Web provides possibility to achieve extensive semantic service, thus ontology library is the core of semantic service. In the view of distributed collaboration system, semantic interconnecting is necessary. The analysis result shows that the main factors causing semantic interconnecting hardly are matching and integration between ontologies. The frame model of integration ontology has been put forward.

Keywords Ontology, Semantic Web, Integration

1 引言

万维网存在两个明显的不足:(1)计算机不能理解网页内容的语义;(2)网上有用信息难找,即使借助功能强大的搜索引擎,查准率也比较低,夹杂了许多用户不需要的信息垃圾。

1998年,在发明万维网10年之后,"万维网之父" Berners. Lee提出了下一代万维网——"Semantic Web"的理念[1]。同年,"网格之父"I. Foster提出的网格计算成为下一代互连网的重要发展方向[2]。结合网格计算、语义网以及互联网服务的优势,弥补各自的不足,将极大地扩充网格的语义能力,并提升语义互联网的计算能力。集成语义服务并完善语义互联是下一代互连网无法回避的问题。目前,语义网和网格技术已成为解决上述问题的最富成果的研究方向。

基于语义服务和Web服务的分布式协作系统是互联网应用系统的发展方向,从某种意义上讲,语义互联、语义共享就是Ontology间的概念匹配和语义匹配的问题,因此,研究本体概念语义的匹配和本体的集成具有重要意义。

2 本体的集成

Ontology是概念模型的明确的规范说明,其核心含义是概念模型(conceptualization)、明确(explicit)、形式化(formal)和共享(share)[4]。语义Web的体系结构如图1所示,Ontology是语义网体系结构中描述语义的核心层,因此,本体是实现语义服务的关键。

本体从本质上讲是领域专家知识的计算机描述,是对该领域概念本质的描述,现在有大量的本体已经开发和正在开发,在构建新的本体时,如何有效地共享和重用已有的领域本体知识已成为本体研究中亟待解决的问题。即使在同一领域开发的不同本体之间,因参加开发的专家的知识难免存在片面性,很难构建一个可以满足所有成员使用需求的本体。那些在应用中较为成功的本体却大都来自于那些绝大多数的专家可以在概念上达成共识的领域。本体集成技术的研究是解决这一问题的有效途径。

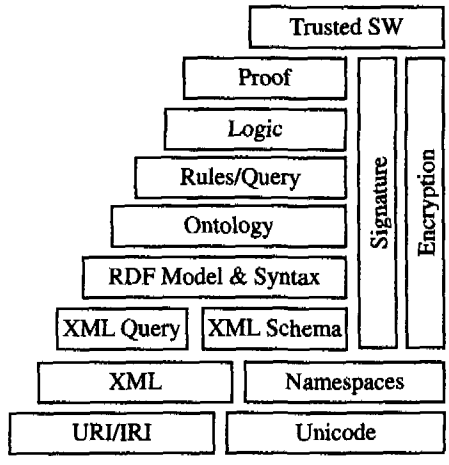

图1 语义Web的体系结构[W3C]

考虑如下的情形(图2,简化的Ontology概念树,连线表示属性关系),客户在互联网中订购旅行机票,各航空公司都

*)CNGI远程教学公用通信平台系统(101048620060012)2005年度国家发改委科学研究计划项目资助。欧 灵 副教授,博士研究生,主要研究方向为语义网、网格计算、网络安全、计算机远程教育;张玉芳 副教授,主要研究方向为计算机远程教育、网格计算;吴中福 教授,博士生导师,主要研究方向为计算机网络与通讯、网络安全、网格计算;钟 将 讲师,主要研究方向为网络安全、免疫计算。

有各自独立建立的票务系统，系统中的概念通过各自的 Ontology 进行解释，而客户通过 Agent 在网上订票时可能按自己的理解发出订购指令，由于概念及语义的不一致，完全可能导致返回的结果与用户的要求不一致，解决问题的办法是实现系统间语义的集成和匹配。

例如：客户提出需要订一张有改签权的往返机票，由于 A 航空公司的 Ontology 没有“返程”概念的明确说明，它用价格和两张单程机票来反应这种需求；B 航空公司的 Ontology 又没有“改签权”概念的明确说明，它用机票价格来反应各种改签权。这样，客户的订票需求得不到满足。

如果各个航空公司能够给用户提供一个统一而全面的语义解释，即集成互联后的 Ontology，用户 Agent 就能通过语义互联网寻找到更多、更准确的服务资源，返回更加满足客户要求的机票信息。

显然，解决类似这样问题的分布式协作系统都需要本体间的互联集成以及基于语义的服务。

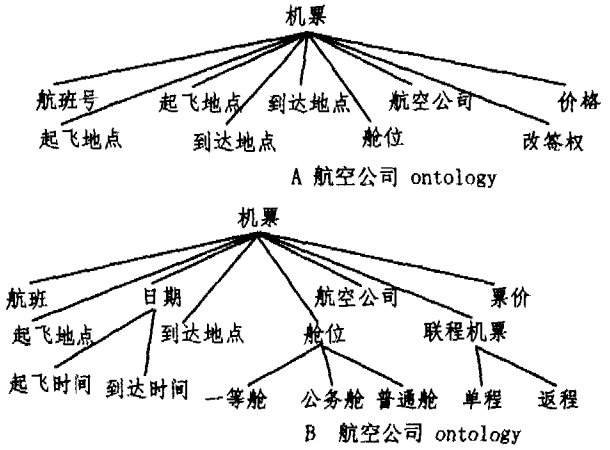

图 2 订票系统本体

Ontology 是语义网研究中的重点和难点，存在大量问题有待解决，如：本体构建、本体评价、本体演化和维护、本体转换和集成等等[8]。

本体库集成的核心问题是本体概念匹配，寻找不同本体间概念、属性、个体之间的共同点和不同点。

现有的系统都是使用集中式的、一致性的、可扩充的 Ontology 库，分布式协作系统、集成系统中的主要问题是本体的互联、协作、集成。造成语义互联困难的因素主要有：1）不同的本体开发语言和系统；2）不同的本体使用不同的概念（词汇）表示同一概念；3）同一概念在不同的本体中表达不同的含义；4）各本体使用不同的结构来表示相同（或相似）的信息；5）各本体中的概念之间存在着各种联系，但因为各信息源的分布自治性，这种隐含的联系不能体现出来。

对于问题 1），成立于 2001 年 2 月的 W3C Semantic Web Activity 工作组展开的 XML、XMLschema、RDF、RDFschema、DAML+OIL、OWL 等标准化工作是近年来语义网研究的重要成果，标准的本体语言为实现本体的互联、协作、集成奠定了更加坚实的基础。

我们注意到现在在同一领域存在着大量不同的本体，它们分别支撑自己的系统的运行。例如，在计算机科学领域的知识表示中就有四种类型的 Ontology 分类[5]。如果能整合这些 Ontology，将极大地方便领域中这些异构系统的互操作，使得机器间、机器与人之间的互信得以实现。虽然开发一个大规模的 Ontology 并使之支持该领域的各种系统也是问题的解决办法，但是，这是一个长期而艰巨的任务。通过对在现有的 Ontology 基础上的整合或互联，实现异构系统间的互操作也是一个可行的解决办法，同时，在此基础上也可以构造更大应用范围的 Ontology。

因此，对于问题 2）、3）、4）、5），我们需要找到一个解决求解本体间相容性、语义一致性的框架。

要实现本体间的精确匹配是一件极具挑战性的工作，已有文献表明：1）可以通过计算机自动发现匹配（极其困难）；2）通过人机协作发现并定义匹配，描述一定条件下的匹配。

本文第 3 部分提出了一个 Ontology 集成的框架模型，第 4 部分给出了一个基于 Agent 的集成 Ontology 的语义 Web 服务模型。

3 集成 Ontology 的框架

斯坦福大学的 Natalya Fridman Noy 等人认为：Ontology 的集成可以分为两类：一类称为合并（Merge）；一类称为联合（Align），可以使用一些规则来定义两个 Ontology 中概念的关系；如图 2 所示。

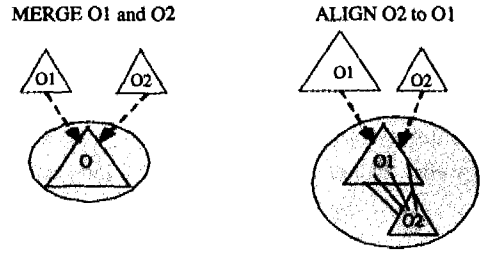

图 3 本体集成的分类

本体集成的核心问题是寻找不同本体系统间的概念语义匹配，目前计算机领域的研究者已经在自动 Ontology 的自动集成、半自动集成以及工具支持下的集成方面展开研究。比较典型的工具有：OntoWeb，Protege PROMPT 等。

我们设计了一个寻找本体映射的模型框架，如图 4。

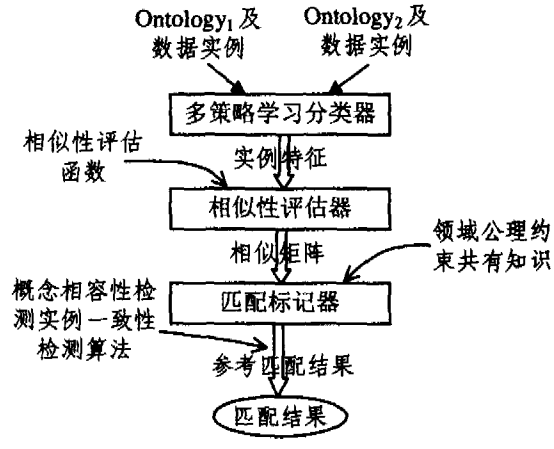

图 4 本体集成的框架模型

获取映射的过程是：

先将 O1 和 O2 的分类概念树及实例送入多策略学习分类器，通过交叉学习分类，利用多策略的学习分类方法获取尽量完整的基于分类树概念的实例特征，并送入相似性评估器。

相似性评估器根据实例特征，在相似度评估策略的指导

（下转第 260 页）

9 Flanagan C, Qadeer S. Predicate abstraction for software verification. The 29th Annual ACM SIGPLAN - SIGACT Symposium on Principles of Programming Languages, Portland, USA, 2002
10 Ball T, Millstein T, Rajamani S K. Polymorphic predicate abstraction. ACM Transactions on Programming Languages and Systems, 2005,27(2): 314～343
11 Ball T, Rajamani S K. Boolean programs: A model and process for software analysis:[Technical Report]. Redmond, USA: Microsoft Research, 2000
12 Das M. Unification-based pointer analysis with directional assignments. The 2000 ACM SIGPLAN Conference on Programming Language Design and Implementation, Vancouver, Canada, 2000
13 Ball T, Rajamani S K. Bebop: A symbolic model checker for Boolean programs. The 7th International SPIN Workshop on Model Checking of Software, Stanford, USA, 2000
14 Ball T, Rajamani S K. Bebop: A path-sensitive interprocedural dataflow engine. The 2001 ACM SIGPLAN-SIGSOFT Workshop on Program Analysis for Software Tools and Engineering, Snowbird, USA, 2001
15 Reps T, Horwitz S, Sagiv M. Precise interprocedural dataflow analysis via graph reachability. The 22nd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, San Francisco, USA, 1995
16 Reps T, Horwitz S, Sagiv M. Precise interprocedural dataflow analysis with applications to constant propagation. Theoretical Computer Science, 1996,167: 131～170
17 Kurshan R P. Computer-aided Verification of Coordinating Processes. Princeton, USA: Princeton University Press, 1994
18 Rusu V, Singerman E. On proving safety properties by integrating static analysis, theorem proving and abstraction. The 5th International Conference on Tools and Algorithms for the Construction and Analysis of Systems, Amsterdam, the Netherlands, 1999
19 Lakhnech Y, Bensalem S, Berezin S, et al. Incremental verification by abstraction. The 7th International Conference on Tools and Algorithms for Construction and Analysis of Systems, Genova, Italy, 2001
20 Ball T, Rajamani S K. Generating abstract explanations of spurious counterexamples in C programs:[Technical Report]. Redmond, USA: Microsoft Research, 2002
21 Henzinger T A, Jhala R, Majumdar R, et al. Abstractions from proofs. The 31st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, Venice, Italy, 2004
22 Craig W. Linear reasoning: A new form of the Herbrand-Gentzen theorem. Journal of Symbolic Logic, 1957, 22(3):250～268
23 Pudlák P. Lower bounds for resolution and cutting plane proofs and monotone computations. Journal of Symbolic Logic, 1997,62(2):981～998
24 Chaki S, Clarke E, Groce A, et al. Modular Verification of Software Components in C. ACM Transactions on Software Engineering,2004,30(6):388～402
25 Flanagan C, Leino K R M, Lillibridge M, et al. Extended static checking for java. The 2002 ACM SIGPLAN Conference on Programming Language Design and Implementation, Berlin, Germany, 2002
26 Flanagan C, Qadeer S. Predicate abstraction for software verification. The 29th Annual ACM SIGPLAN - SIGACT Symposium on Principles of Programming Languages, Portland, USA, 2002
27 Havelund K, Pressburger T. Model checking Java programs using Java PathFinder. International Journal on Software Tools for Technology Transfer, 2000,2(4):366～381
28 Corbett J, Dwyer M, Hatcliff J, et al. Bandera: Extracting finite-state models from Java source code. The 22nd International Conference on Software Engineering, Limerick, Ireland, 2000
29 Dwyer M, Hatcliff J, Joehanes R, et al. Tool-supported program abstraction for finite-state verification. The 23nd International Conference on Software Engineering, Toronto, Canada, 2001

(上接第 188 页)

下,发现基于实例特征的位于不同 Ontology 分类树的两个概念对的相似性,同时判断一个概念在另一个 ontology 中的相对位置,获取基于实例特征的概念相似性矩阵及位置特征矩阵。

匹配标记器(如:Hidden Markov Model Labeler)在分析树结构、领域公理约束、共有知识的基础上,利用概念相似性矩阵及位置特征矩阵标记映射关系,获得参考匹配结果。

利用概念相容性检测、实例一致性检测推理机(如:Tableau Algorithms)检测参考匹配结果,最后获得可以信赖的结果。当然,半自动方式下进行的有条件映射标记,也是可以接受的(如:Protege PROMPT)。

总结 语义服务是下一代 WebWeb 服务必须解决的难点问题,语义网为实现广泛的语义服务提供了可能,Ontology 是语义网体系结构的核心。分布式协作系统的重要问题是本体的匹配和集成,针对协作的分布式系统需要语义互联的问题,本文分析了造成语义互联困难的主要因素是本体的匹配和集成,提出了一个基于机器学习的 Ontology 集成的框架模型。该框架的四个关键步骤是:学习分类算法及策略、相似性评估、标记算法、相容性一致性检测。

基于机器学习的本体概念匹配研究才刚刚起步,本文提出的这个框架,结构完整,每一步工作的目标明确,我们将在之后的研究论文中逐一讨论。

参 考 文 献

1 Berners-Lee T, Hendler J, Lassila O. The Semantic Web, Scientific American, 2001,284(5):34～43
2 Foster I, Kesselman C, Nick J M, et al. The Physiology of the Grid - An Open Grid Services Architecture for Distributed Systems Integration. http://www.globus.org/ogsa/, Feb. 2002
3 Noy F N. What Do We Need for Ontology Integration on the Semantic Web, Position Statement. In: Proc. of the Workshop on Semantic Integration, the 2nd International Semantic Web Conference, Sanibal Island, Florida, USA, Oct. 2003
4 Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods. Data and Knowledge Engineering, 1998,25(122):161～197
5 Jurisica I, Mylopoulos J, Yu E. Ontologies for Knowledge Management: An Information Systems Perspective. Knowledge and Information Systems, 2004,6(4):380～401
6 Baader F, McGuinness D, Nardi D, Schneider P P. The description logic handbook: theory, implementation and applications [Z]. Cambridge: Cambridge University Press, 2002
7 Naing Myo-Myo, Limy Ee-Peng, Hoe-Lian Dion Goh. Ontology-based Web Annotation Framework for HyperLink Structures. In: Proceedings of the Third International Conference on Web Information Systems Engineering (WISE), Singapore, 2002
8 李善平,尹奇,胡玉杰,等. 本体论研究综述. 计算机研究与发展, 2004,41(7)