语义网环境下的信息资源整合模式

章成志，苏新宁，邓三鸿

（南京大学 信息管理系，江苏 南京 210093）

摘 要：语义网是当前WWW的扩展，其目标是帮助人类和计算机更好地协同工作。作者分析了语义网环境下信息资源整合的背景、特点及体系结构，本体在数据层、元数据层和表现层等三个信息资源整合层面上的应用模式，指出亟待解决的问题。

关键词：信息资源整合；语义网；本体；元数据

中图分类号：G203　　**文献标识码**：A　　**文章编号**：1007-7634（2006）09-1335-04

Information Resources Integration Mode under the Semantic Web

ZHANG Cheng - zhi, *SU Xin - ning*, *DENG San - hong*

(*Department of Information Management*, *Nanjing University*, *Nanjing* 210093, *China*)

Abstract: Semantic web is expansion of WWW and it aims at promoting human cooperating with computer. This paper expounds the background, characteristic and architecture of information resource integration for the semantic web. Information resource integration methods from three levels, e. g. data level, metadata level and presentation level are provided. The key issue and difficulty of the information resource integration for the semantic web is put forward.

Key words: information resources integration; semantic web; ontology; metadata

1 引　　言

随着互联网日益普及与深入应用，信息资源整合这一研究热点从异构数据集成逐渐转变为对Web信息资源的整合[1]。20世纪90年代后期，随着人工智能和知识工程研究领域对本体工程研究的不断深入，基于本体（Ontology）的信息资源整合方法和技术成为研究热点[2]，信息资源整合研究逐渐转变为语义网（Semantic Web）环境下的信息资源整合研究[3-4]（Information Integration for Semantic Web，以下简称SWII）。语义网环境下的信息资源整合更侧重于Web上的智能应用，通过带有语义信息的标注，使得机器能达到一定程度的智能，解决传统的Web所无法解决“信息爆炸，知识贫乏”问题。

本文分析了语义网环境下信息资源整合的背景、特点及体系结构，本体在数据层、元数据层和表现层等三个信息资源整合层面上的具体应用，详细说明了基于本体的信息资源整合方法中的若干关键技术，指出亟待解决的问题。

2 SWII背景与特点

2.1 SWII的背景

语义网为信息处理提供了新的机遇和挑战。语

收稿日期：2005-12-20

作者简介：章成志(1977-)，男，安徽望江人，博士研究生，从事信息检索、信息集成与中文信息处理研究；苏新宁(1955-)，男，江苏南京人，教授，博士生导师，从事网络信息资源的研究与开发、情报检索算法、中文信息处理技术研究；邓三鸿(1975-)，男，湖北郧州人，博士，讲师，从事知识管理研究.

义Web的目标是使得Web上的信息具有计算机可以理解的语义，满足智能软件代理（Agent）对WWW上异构和分布信息的有效访问和检索。语义网各层描述如表1所示。

表1 语义Web体系结构

层数	名称	描 述
第一层	UNICODE和URI	Unicode用于资源的统一编码，URI用于标识资源
第二层	XML + NS + xml schema	用于表示数据的内容和结构
第三层	RDF + rdf schema	用于描述Web上的资源及其类型
第四层	Ontology vocabulary	用于描述各种资源之间的联系
第五层	Logic	
第六层	Proof	在上面四层的基础上进行的逻辑推理操作
第七层	Trust	

注：语义网基本体系结构图参见：http://www.w3.org/2001/Talks/0228-tbl/slide5-0.html

表1中，第二至第四层为语义网体系结构的核心层。XML和RDF都能为所表述的资源提供一定的语义。但是XML中的标签（tags）和RDF中的属性（properties）集都没有任何限制。XML和RDF在处理语义上存在的问题是：同义词及多义词问题。本体是概念化的明确的规范说明[5]。在语义网中，本体具有非常重要的地位，是解决语义层次上Web信息共享和交换的基础。Ontology通过对概念的严格定义和概念之间的关系来确定概念精确含义，表示共同认可的、可共享的知识，从而解决上面的问题。目前广泛使用的Ontology有Wordnet（http://wordnet.princeton.edu）、CYC（http://www.cyc.com）、Framenet（http://framenet.icsi.berkeley.edu）、SENSUS（http://mozart.isi.edu:8003/sensus）等。为了便于Web上应用程序使用方便，需要有一个通用的标准语言来表示本体，就像XML作为标准的数据交换语言一样。在本体描述语言方面，Ontolingua[6]、LOOM[6]等传统的本体描述语言已经在基于知识的应用中被用于知识表达。随着Web技术的发展，相继出现很多在WWW环境下的本体描述语言：RDF、RDF Schema、SHOE（http://www.cs.umd.edu/projects/plus/SHOE）、XOL（http://www.ai.sri.com/ pkarp/xol）、OIL（http://www.ontoknowledge.org/oil）、DAML（http://www.daml.org）+ OIL以及基于RDF(S)和DAML + OIL的Web本体语言OWL（http://www.w3.org）。

SWII正是利用本体，依据语义网具有理解语义，能够推理的这一功能对分布于网络上的异构信息资源进行语义层次的集成与整合。从信息处理的智能化，处理资源的复杂化等角度来看，语义网为信息资源的整合提出了新的挑战和机遇。

2.2 SWII的特点

语义网环境下的信息资源整合不同于传统的数据库数据集成和传统Web信息集成的地方就在于，SWII具有整合方法的智能性和广泛适用性。

（1）整合方法的智能性。语义网引入本体技术及逻辑推理操作可解决信息源的结构异构及语义异构问题。SWII引入了人工智能和知识工程中的大量智能方法及专家知识，经过本体学习、本体标注对WWW上的信息资源进行智能化的信息处理，如基于本体的聚类、分类，文本可视化等。SWII的目标是信息资源懂得动态“自组织”，以达到“自整合”功能。

（2）整合方法的广泛适用性。语义网是对现有Web技术的扩展，其中的很多方法和技术，如本体技术可应用于信息系统，只要有元数据的地方，就可以用RDF，因为RDF本身就是一种元数据表示语言。在数据库系统中，元数据是库表的Schema，在XML信息系统中，元数据是XML的标签和XML Schema，而RDF + RDF Schema完全可以作为它们的替代品，而它的优势之一就是它是有语义的，机器可理解的，而且Web化。

3 SWII结构体系

将传统的信息资源整合体系结构与语义网的应用特色结合起来，便得到SWII的结构体系。SWII分别从三个层面，即数据层、元数据层、表现层，借助与语义网中本体技术及逻辑推理功能，进行基于语义的信息资源整合。下面对这三层的信息资源整合方法分别加以说明。

3.1 数据层

数据层整合是一种物理整合方法，是在“数据大集中”思想指导下进行的信息资源整合方法。数据层整合是对现有的信息资源重新组织、深度加工

和知识服务的过程。它需要建立一个新的存储仓库，将收集到的各种资源装入其中，不同结构的信息资源被组织为相同的数据格式，用统一的检索平台很容易的检索所有的信息资源。数据层整合的优势在于，经过该层整合后，便于进行数据的统一存储或迁移，便于在其上进行数据挖掘，进行分析和决策等。

SWII中的数据整合是一种基于本体的，对数据进行语义整合的信息整合方法，即基于本体的数据整合。举例来说，比如一本书籍由大陆学者和台湾学者共同编写，在名词的用法上会有不同，编辑在最后校编的时候，就要把这些不同的用法统一起来。若此过程是借助于本体自动完成的，那么这就是一种典型的、较简单的基于本体的数据整合过程。

基于本体的数据集成的内容主要包括：信息的格式转换，基于全文的自动标引、自动分类，信息过滤与去重等。本体的引入，使得这些自动化加工处理从传统的语法层次上升到语义层次。图1～图2为一个基于本体的自动标引的例子，其中图1表示基于本体的自动标引过程，主要包括如下几个步骤。①Web信息提取，对Web页面进行信息提取，获取关键词。②利用某一领域本体对关键词进行语义关联。③完成Web的语义标注，给出语义可视化图形或。④给出自动标引的结果，如图2所示。

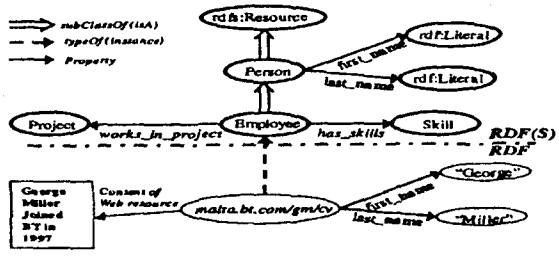

图1　一个基于本体的自动标引例子[7]

Descriptor	Class	Property	Resource
Miller	Employee	Ø	malta. bt. com/gm/cv
joined	Employee	Ø	malta. bt. com/gm/cv
BT	Employee	Ø	malta. bt. com/gm/cv
1990	Employee	Ø	malta. bt. com/gm/cv
George	Employee	first _ name	malta. bt. com/gm/cv
Miller	Employee	last _ name	malta. bt. com/gm/cv

图2　自动标引的结果[7]

3.2　元数据层

元数据（metadata）是关于数据的组织、数据域及其关系的信息。元数据为各种形态的数字化信息单元和资源集合提供规范、一般性的描述。资源描述框架（RDF）是一种元数据，它用XML作为交换语法，提供应用之间的互操作，这种框架对Web资源进行描述，方便信息的自动处理。

图3给出了在元数据层上进行信息资源整合的流程，主要步骤如下。①预处理。主要是剔除在格式、内容、语言等方面存在的问题或严重缺失的文档，产生格式相对规整的文本文档。②提取元数据。由数字化文档元数据的规范定义，产生提取元数据的各种模式，依据元数据模式进行数字化文档的挖掘与匹配，利用与数字化文档相关的启发式规则和经验规则，有效提取元数据。③元数据的语义关联与逻辑推理。借助本体对元数据进行语义层次上的关联，如文献的主题之间，来源文献与引文文献之间的关联等都可以利用本体进行语义关联或逻辑推理。

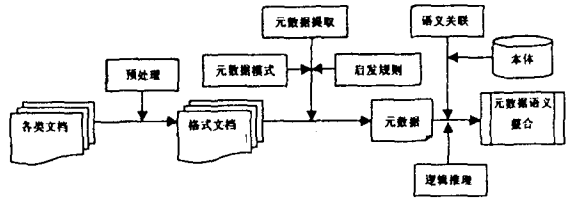

图3　元数据层整合流程

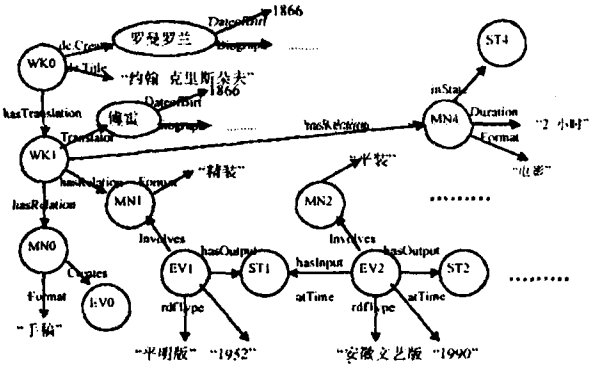

图4　用本体表述傅雷翻译作品《约翰·克里斯朵夫》[8]

图4为一个使用本体进行元数据关联的例子。

3.3　表现层

表现层整合是一种逻辑整合方法，它是将各种应用按照语义进行集成，建立语义导航或语义门户等，为用户提供基于语义的整合服务。所谓导航，是指这样一种信息浏览机制，即：用户在Web上浏览信息时，通过站点的目录、站点地图或站内搜

索等方法搜寻他们所需要的信息。语义网中的导航被称为语义导航（Semantic Navigation）。Hp 的 Semantic Blogging（http: //jena.hpl.hp.com/~ stecay/papers/xmleurope2004/040420 _ semblog _ draft10.html # semnav），其基本机制为语义导航。语义门户（Semantic Portal）主要实现了语义浏览，即按目录语义检索，就像我们常用的 MIS 系统那样，可以按某些字段检索。语义门户方面的典型案例有：OntoWeb（http: //ontoweb.ontoware.org），Knowledge Web（http: //knowledgeweb.semanticweb.org）等。图 5 给出 Hp 实验室语义门户总体框架结构。

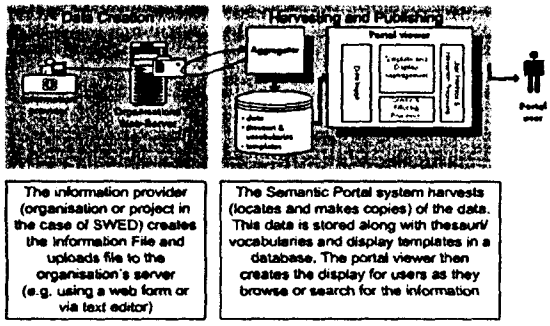

图 5 Hp 实验室语义门户总体框架结构图[9]

4 结 语

语义网是当前 Web 技术的扩展。在语义网环境下，通过带有语义信息的标注，使得机器能达到一定程度的智能，解决传统的 Web 所无法解决“信息爆炸，知识贫乏”问题。语义网环境下的信息资源整合是一种基于语义的信息集成，为用户提供基于语义的服务，大大提高了信息加工和服务的知识含量，为知识创新提供了坚实的技术。语义网环境下的信息资源整合需要数据库界、人工智能界、图书情报界以及各领域的专家共同努力，相互配合，协同进行研究与开发。本文旨在提出研究语义网环境下信息资源整合的基本框架，其中的关键技术及应用有待进一步深入研究。

参考文献

1 May Wolfgang, Lausen Georg. A Uniform Framework for Integration of Information from the Web[J]. Information Systems. 2004, 29(1): 59 - 91.

2 Martin Doerr, Jane Hunter et. al. Towards A Ccore Ontology for Information Integration[EB/OL]. http://jodi. ecs. soton. ac. uk/Articles/v04/i01/Doerr, Accessed: Aug. 1, 2005 - 11 - 05.

3 Holger Wache, Thomas Vogele et. al. Ontology - based Integration of Information - ASurvey of Existing Approaches[J]. In IJCAI - 01 Workshop: Ontologies and Information Sharing, Eds., Asuncion Gomez Perez, Michael Gruninger, Heiner Stuckenschmidt, and Mike Uschold, Seattle, WA, 2001, (8): 108 - 117.

4 Ubbo Visse, Gerhard Schuster. Finding and Integration of Information - A Practical Solution for the Semantic Web[J]. In: Proceedings of ECAI 02, Workshop on Ontologies and Semantic Interoperability, Eds., Euzenat, J., Gomez - Perez, A., Guarino, N. and Stuckenschmidt, H., Lyon, France, 2002, (7): 73 - 78.

5 Tom Gruber. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993, 5(2): 199 - 220.

6 邓志鸿，唐世渭，等. Ontology 研究综述[J]. 北京大学学报（自然科学版），2002, 38(5): 730 - 738.

7 John Davies and Frank van Harmelen. Towards the Semantic Web: Ontology - driven Knowledge Management[M]. John Wiley And Sons, 2003. 135 - 138.

8 刘 炜，李大玲，夏翠娟. 基于本体的元数据应用[EB/OL]. http://eprints. rclis. org/archive/00003403/01/ontology _ basedMedadata2. pdf, Accessed: Aug. 1, 2005 - 11 - 05.

9 Paul Shabajee[EB/OL]. http://www. swed. org. uk/swed/about/swed _ approach. htm, Accessed: Aug. 1, 2005 - 11 - 05

（责任编辑：徐 波）