基于多 Agent 的语义 Web 挖掘系统模型研究

王涛伟[1]，任一波[2]
（1.浙江万里学院，宁波 315100；2.浙江工商职业技术学院，杭州 315012）

摘 要：随着语义 Web 技术的出现，传统的 Web 挖掘面临新的挑战. 文章讨论了语义 Web 的体系结构、Web 挖掘和 Agent 技术的有关概念，提出了基于多 Agent 的语义 Web 挖掘系统模型，并对其进行了分析. 实验结果表明，该系统模型具有较高的准确性和效率.

关 键 词：语义 Web；Web 挖掘；多 Agent
中图分类号：TP393 **文献标识码：**A **文章编号：**1671－2250（2006）05－0041－04
收 稿 日 期：2006－04－14

作 者 简 介：王涛伟，浙江万里学院计算机与信息学院讲师；任一波，浙江工商职业技术学院计算机系讲师.

现今 Web 上的信息基本都是用 HTML 或者 XML 来表示的，通过浏览器人们可以很直观地获得网页中的信息，但是随着网上信息量的迅速膨胀，其简单的结构很大程度上限制了 Web 的高级应用，如信息检索，Web 挖掘等. 为了解决这个困难，Tim Berners-Lee 提出了 Semantic Web(语义 Web)的概念[1]，有人称之为下一代的网络．本文介绍了语义 Web 和语义 Web 挖掘的相关知识，根据其特点，结合 Agent 技术提出了一个基于多 Agent 的语义 Web 挖掘系统模型结构．

1 语义 Web 和语义 Web 挖掘的有关概念

1．1 语义 Web 的体系结构

语义 Web 就是对现有网络的一种扩展，在这个扩展层上，信息被赋予了语义，使得机器可以通过这些语义更好地发掘 WWW 信息的潜力. Tim Berners-Lee 一直致力于语义 Web 技术的研究，并一直关注语义 Web 技术的发展，在综合了语义 Web 研究领域的最新成果的础上，提出了语义 Web 模型[2]，他把语义 Web 描述为七层结构，如表 1 所示. 从中可以看出他所建议的语义 Web 层次结构. 语义 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．

表 1 语义 Web 七层结构

层数	名称	描述
第一层	Unicode+URI	语义 Web 网络的基础，Unicode 处理资源的编码，URI 标识资源
第二层	XML+NS+xml Schema	用于表示资源的内容和结构
第三层	RDF+rdf Schema	用于描述 Web 上的资源及其类型
第四层	Ontology vocabulary	本体，用于描述各种资源之间的联系
第五层	Logic	进行逻辑推理操作
第六层	Proof	
第七层	Trust	

第一层是整个语义网络的基础，Unicode 处理资源的编码，URI 负责标识资源. 第二层用于表示数据的内容和结构，他使用 XML 进行文档结构化，使用 XML Schema 定义 XML 文档的结构约束. NS 是名字空间. 第三层用于描述 Web 上的资源及其类型，RDF 描述对象(或者资源)以及它们之间关系，他为数据模型提供了简单的语义，这些数据模型能够用 XML 语法进行表达. RDF Schema 用于描述 RDF 资源的属性和类型的词汇表，提供对这些属性和类型的普遍层次的语义. 第五、六、七层是在上述层的基础上进行的逻辑推理操作.

根据该体系结构，语义网的实现离不开 XML (可扩展标记语言 eXtensible Markup Langauge)和 RDF（资源描述框架 Resource Description Framework）. XML 于 1998 年由 World Wide Web Consortium(W3C)设计出来，是一种用于定义标记语言的工具，其内容包括 XML 声明、用以定义语言语法的 DTD(document type declaration 文档类型定义)、描述标记的详细说明以及文档本身，而文档本身又包含有标记和内容. RDF 是 W3C 新建的标准，它通过描述对象[O]-属性 A-值[V]三元组形式或 A(O ,V)形式的关系体现相关事务的信息内容，对 Web 信息内容进行语义化的描述. RDF 定义了一个简单的模型用来表示 RDF 的元数据，该模型通过命名的属性和属性值来表示资源之间的关系和资源内部的关系，如图 1 所示表示：Web 页面“ http：//www. w3. org/employee/ zh ”有一个名字“ 张三 ”，它是书“ http：// www. books. org/ISBN1002 “ 的作者. 这本书的价格是 45 元.

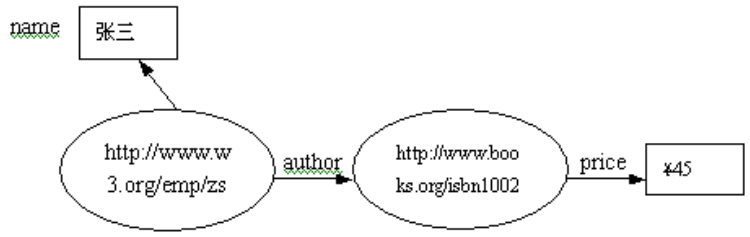

图1 资源之间的图表示形式

RDF Schema是用来描述RDF中所用到的属性和资源之间关系的语言，他们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，如同一概念有多种词汇表示，同一词汇有多种含义. 因此，必须在语义层次上解决Web信息共享和交换的问题.

Ontology(译为本体)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，因此，在语义 Web 中 Ontology 非常重要，他是解决上述问题的基础，为结构的第四层. 在使用 XML 定义标签格式和 RDF 表达数据后，使用一种 Ontology 的网络语言来描述网络文档中的术语的明确定义及其之间的关系. 目前，主流语义 Web 语言有 OIL (Ontology Inference Layer)、DAML+OIL (即 DARPA 代理标记语言+本体推论语言)和 OWL (Web Ontology Langauge，Web 本体语言). 其中 OIL 是一种本体描述语言，它建立在 RDF 和 RDF Schema 之上，为其提供概念的定义以及逻辑描述，OIL 结合了三个方面的内容：描述逻辑中的形式语义和推理支持、基于框架系统中的认知建模原语和网络语言的基于 XML 和 RDF 的语法，DAML+OIL 它比 OIL 提供了更丰富的逻辑描述. OWL 是 DAML+OIL Web 本体语言的修改版，并吸取了在 DAML+OIL 的设计和应用中得到的经验教训.

1．2 Web 挖掘

Web 挖掘它是指从 World Wide Web 上发现、提取感兴趣的有用模式和隐含的、事先未知的、潜在的信息. 按照处理对象的不同，一般将 Web 挖掘分为类 Web 三类：Web 内容挖掘、Web 结构挖掘和 Web 使用记录挖掘.

(1)Web 内容挖掘 Web 内容挖掘是从 Web 文档内容或其描述中抽取知识的过程. Web 文档文本内容的挖掘、基于概念索引的资源发现以及基于代理的技术都属于这一类. Web 内容挖掘又分为文本挖掘和多媒体挖掘，对象分别是 Web 文本信息和 Web 多媒体信息.

(2) Web 结构挖掘 Web 结构挖掘是从 Web 站点结构和 Web 页面结构中推导出新的模式和知识，因此结构挖掘的重点在于链接信息. 它不仅仅局限于文档之间的超链接结构，还包括文档内部的结构、文档 URL 中的目录路径的结构等. 由于文档之间存在关联，使 WWW 能够提供文档内容之外的有用信息. 利用这些信息，可以对页面进行排序，发现重要的页面.

(3)Web 使用记录挖掘 Web 使用记录挖掘是从 Web 的访问记录中抽取感兴趣的模式. WWW 中的每个服务器都保留了访问日志，记录关于用户访问和交互的信息. 分析这些数据可以帮助理解用户的行为，从而改进站点的结构，或者为用户提供个性化的服务.

1．3 语义 Web 挖掘

语义 Web 挖掘旨在将 Web 挖掘和语义 Web 这两大研究领域结合起来，使它们相互促进，共同发展. 因此一般将语义 Web 挖掘也分为语义 Web 内容挖掘、语义 Web 结构挖掘和语义 Web 使用挖掘三种. Web

挖掘的结果有助于构建语义 Web，语义 Web 的语义知识使得 Web 挖掘更易实现，且能改善 Web 挖掘的结果. 另外在语义 Web 上进行挖掘与传统的 Web 挖掘过程也有很大的差别.

2　Agent 及多 Agent 系统的基本概念

Agent 是一个能在特定环境下连续自发地实现功能，同时与相关 Agent 和进程联系的软件实体. Agent 通常具有以下基本特征[3]：

自主性：能够在没有人或其他 Agent 干预下完成大部分功能，控制其内部状态.

社会能力：能够和其他 Agent 或人进行交互，实现其目标.

被动响应能力：能够感知周围环境的变化，并产生实时响应.

主动响应能力：能够主动地根据自身目标进行活动.

自适应性：能够通过自身的学习机制来适应环境.

可移动性：可以携带数据和指令移动到其他环境中并在那里执行指令.

多 Agent 系统是指由多个 Agent 组成的处理分布式问题的系统，它通过各个 Agent 之间的交互和协作来实现系统的整体目标. 它不仅具有单 Agent 的许多特征，而且还能够解决单 Agent 不能解决的复杂问题. 多 Agent 体现了单个 Agent 的自治能力和各个 Agent 之间的协同能力两方面. 多 Agent 具有高度的开放性、灵活性、广泛的适应性和简单的设计性，被认为是下一代软件开发的新标准.

3　基于多 Agent 的语义 Web 挖掘系统模型结构及工作机制

3．1　系统模型的结构和各 Agent 的主要功能

根据语义 Web 和 Web 挖掘的特点，结合多 Agent 技术，提出了基于多 Agent 的语义 Web 挖掘系统框架模型. 系统由以下这些 Agent：资源搜集 Agent、RDF 数据学习 Agent、算法挖掘 Agent、本体 Agent、决策 Agent、人机界面 Agent 以及 RDF 数据库和本体库系统组成如图 2 所示．具体功能如下：

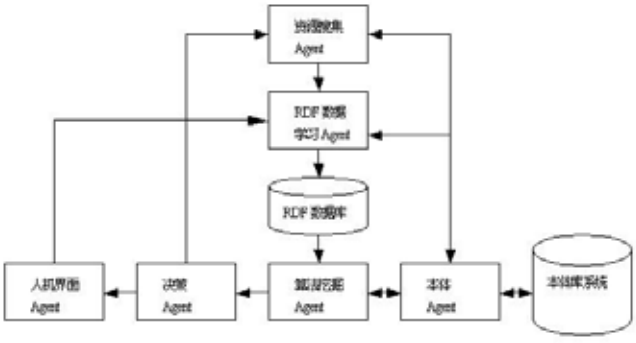

图 2　多 Agent 语义 Web 挖掘模型结构图

(1)资源搜集 Agent　资源搜集 Agent 的主要任务是搜集任务相关的数据集，并且根据决策 Agent 的要求调整资源搜集，最后将资源搜集的结果提供给 RDF 学习 Agent 进行本体学习. 这里资源搜集 Agent 和本体 Agent 之间有相互协作工作：即通过本体 Agent 和本体库系统交互取得要进行搜集的资源的语义信息，进行基于语义的资源搜集. 同时资源搜集 Agent 对资源搜集中发现的新的本体描述可以通过本体 Agent 将它们添加到本体库系统中.

(2)RDF 学习 Agent　通过对资源搜集 Agent 搜集到的数据进行本体学习．如通过聚类方法，可将最相关的数据聚集在一起并且建立索引，可以大大地减少下一步知识发现时的数据预处理工作量. 这里也有一个和本体 Agent 的交互，因为在对 RDF 数据进行本体学习时，必须通过本体 Agent 和本体库系统交互来取得各个资源描述的语义信息.

(3)RDF 数据库　根据 RDF 学习的结果，把最有关的数据存储在一起. 在这里基本的数据单元是 RDF 三元组，即资源、属性和属性值的形式.

(4)算法挖掘 Agent　它执行知识发现的任务，从 RDF 数据描述中发现潜在的、未知的、使人感兴趣的知识. 与通常的数据挖掘模块不同的是，这里需要通过本体 Agent 利用本体库中本体知识，另外语义 Web 挖掘模块发现的知识可以更进一步的完善本体库中的知识.

(5)决策 Agent　决策 Agent 按照一定的策略，根据挖掘 Agent 所发现的知识，通过人机界面向用户提供服务，同时决策 Agent 也可以向资源搜集 Agent 提供指导.

(6)人机界面 Agent 人机界面是连接系统和用户的桥梁. 一方面人机界面 Agent 将决策后的信息以可视化的方式展示在用户面前，另一方面也可以将用户输入的信息作为学习的数据.

(7)本体 Agent 本体 Agent 实现资源搜集 Agent，RDF 学习 Agent 以及算法挖掘 Agent 与本体库系统的交互. 它主要是提供这些模块中所用到的语义信息查询，同时也可利用这些模块新发现的本体知识来更新本体知识库中的本体知识. 例如，如果在资源搜集的时候发现某一个资源描述使用了一个本体库中所没有存储的本体知识，就需要通过本体 Agent 与本体库系统交互来将这个本体知识加入到本体库系统中. 还有对于语义 Web 挖掘模块新发现的本体知识也要使用本体 Agent 来把它添加到本体库中. 在本系统中本体 Agent 的作用是非常重要的，它是实现资源搜集 Agent，RDF 学习 Agent 以及算法挖掘 Agent 与本体库系统的交互的关键.

(8)本体库系统 在本模型中本体库系统同样非常重要，正是此模块向机器提供了关于 RDF 数据的概念和关系定义、语义描述及基于其上的一系列推理，使得语义 Web 上的信息能为机器所“理解”. 本体库系统是一种分组和重组各种本体以便于更进一步重用、集成、维护、映射和版本化的重要工具. 有利于本体的重用和共享是一个本体库系统最重要的要求，一个本体库系统首先必须支持开放存储、标识和版本化，提供使本体适应某种领域和任务的支持. 它通过本体 Agent 对资源搜集 Agent， RDF 学习 Agent 以及算法挖掘 Agent 提供语义支持. 同时对于这些 Agent 新发现的本体知识进行存储、维护和管理.

3．2 系统模型中各 Agent 的工作机制

从上可见，本系统模型中的各 Agent 本身都具有自治性，各个 Agen 之间具有协作性. 每个 Agent 都是自主性的实体，其内部都有推理机进行知识的推理，所有 Agent 内部都有一个专门的通信模块，用于和其他 Agent 进行消息传递和交换，采用 DARPA 提出的知识查询与操作语言 KQML 作为各 Agent 之间的通信语言. KQML 语言是一种基于消息的 Agent 通信语言，可以根据需要进行动态扩展，是目前一种较为成熟的 Agent 通信语言. 另外模型采用合同网协议[4]作为 Agent 之间的协作机制. 它最早是由 Davis 和 Smith 等人在研究分布式问题求解时提出来的，现在被广泛的应用到多 Agent 系统的协作中.

4 小结

本系统在多 Agent 系统平台上构建了基于语义 Web 的挖掘系统，实现了语义 Web 挖掘的整个过程，并且较好的解决了多 Agent 的通信和协作问题. 为了测试系统的性能和效率，系统利用学院网站 2005 年 4 月 8 日至 4 月 10 日的网站访问信息作为数据进行了测试，利用关联规则算法进行挖掘，表 2 显示了基于多 Agent 挖掘与传统的 Web 挖掘过程系统的执行时间（s）对比. 本文提出了一个基于多 Agent 的语义 Web 挖掘模型，并且进行了分析，由于语义 Web 中有着丰富的语义信息，因此基于多 Agent 的语义 Web 挖掘比传统的 Web 挖掘系统有更高的准确性和效率，但是系统中的某些功能还不够完善，下一步的工作将重点在探索模型中 Agent 之间更有效的协作模式和系统中的机器学习能力.

表 2 基于多 Agent 挖掘与传统 Web 挖掘执行时间对比

记录数	多Agent挖掘	传统的Web挖掘
6000	22	28
12000	39	47
28000	68	112

参考文献：

[1]Tim Berners-Lee,James Hendler and Ora Lassila.The Semantic Web[J].Scientific American,2001,(5).

[2]TimBerners-Lee:SemanticWeb–Architechture[EB/OL].http://www.w3.org/2000/talks/1206-xml2k-tbl/slide10.html.

[3]范玉顺,曹军威. 多代理系统理论、方法与应用[M].北京:清华大学出版社, 2002.

[4] Smith R,Davis R.Frameworks for Cooperation in Distributed Problem Solving[J].IEEE Transactions on Systems,Man and Cybernetics, 1981,11(1):61-70.

（下转第80页）

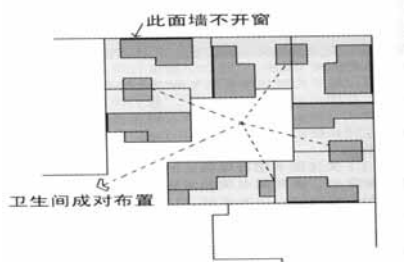

图 3 七个住宅单元组成一个基本组团，沿一个 8m×8m 的小院子周边布置

化等等。这个案例所具备的许多特点，如发展中国家的经济状况，丰富的历史文化底蕴及乡村大范围的居住条件的改善等,都对我国皖南民居及其他人居环境的改善和发展有着很好的借鉴意义。

参考文献：

[1]韩冬青.类型与乡土建筑环境——谈皖南村落的环境理解[J].建筑学报,1993,(8):53-55.
[2]王其钧.传统城镇与民居美学[A].中国传统民居与文化(三)[C].北京:建筑工业出版社,1999:18.
[3]王澍.皖南村镇巷道的内结构解析[J].建筑师,1989,(6):65-67.
[4]汪芳.乡村理想与住宅生长[J].华中建筑,2003,(2):56-58.

Application of SAR Theory in Future Village Buildings: the Hereditary Dwelling Environment in Wannan as an Example

WANG Zhe
(Zhejiang Wanli University，Ningbo 315100)

Abstract：This paper analyzes the dwelling environment of local dwelling in Wannan, subdividing the dwelling space into principal and interest space. Also, the paper discusses the corresponding relationship between “supporting”, “detachable” units originating from the SAR theory and the principal, temperament space in Wannan vernacular buildings, which is expected to be of help for building the future villages into easier dwelling environment.

Key words：local dwelling in Wannan；SAR theory；principal space；temperament space

（上接第 44 页）

Research on the System Model of Semantic Web Mining Based on Multi-agent

WANG Tao-wei[1],REN Yi-bo[2]
(1. Zhejiang Wanli University, Ningbo 315100； 2. Zhejiang Business Technology Institute，Hangzhou 315012)

Abstract：With the appearance of semantic Web technology,traditional Web mining faces some new challenges.This paper discusses some concepts about the system structure of semantic Web,Web mining and agent techniques , presents the system model of semantic Web mining based on multi-agent and analyses them. The experimental results show the model has high accuracy and efficiency.

Key words：semantic Web；Web mining；multi-agent