一个自学习本体支持的辅助学习系统

李向阳

(华侨大学 工商学院信息系,福建 泉州 362021)

摘 要:介绍一个智能辅助学习平台的设计理念、功能和结构。它根据相同领域的电子资料集合自动学习领域本体,基于所学本体自动建立不同资料之间的语义关联,应用语义关联自动实现学习资料之间的参考关系。其结果是将相同领域的原来相互独立的多种资料自动按语义融合为一个整体,以提高学习和研究工作的资料应用效率。

关键词:领域本体;电子学习;本体学习

中图分类号:TP39 **文献标识码**:A **文章编号**:1674-7720(2010)13-0012-03

A learning assisting system supported self-learned ontology

LI Xiang Yang

(Information Management Department, College of Business Administration, Huaqiao University, Quanzhou 362021, China)

Abstract: This paper introduces the designing idea, function and structure of a learning assisting system. It is based on automatic learned domain ontology to establish the semantic relationship among different learning resources, and using the semantic relationship to reference learning materials automatically. As a result, different learning resources integerated semanticly by the system as a whole, material using efficiency in learning and research work enhanced.

Key words: domain ontology; e-learning; ontology learning

目前在网上和电子图书馆已有海量的知识资源,而如何更高效地应用这些电子知识资源已成为一个热门研究和应用领域。与传统纸质知识资源相比,电子知识资源几乎不受容量的限制,能够用计算机对它们进行快速访问和处理。

在学习一门新知识时通常需要对一个主题参考多种资料才能够透彻地理解,在研究工作中,甚至需要分析能够找到的全部知识资源。对纸质知识进行这种资料的查询和引用是一个费时费力、效率低下的工作。本文介绍一个智能辅助学习工具,它帮助用户在学习一份资料的某个知识点时能够自动快速定位其他资料的相同或相近知识点,省去学习时查找资料的时间。

1 系统基本设计理念与功能概述

本节对文中系统用到的本体(ontology)知识点、知识项和等核心概念进行定义,并阐明本设计的基本理念,并简要介绍系统实现的核心功能。

1.1 系统的基本概念和设计理念

(1)本体

本体是知识工程领域一个非常重要的概念,它源于哲学的本体论(ontology),在人工智能中被借用过来表示特定领域知识体系中的概念体系。许多研究者对它有不同的定义,得到公认的是 Tom Gruber 定义:本体是关于共享概念的协议。

本体在实际应用中表现为特定领域中专业术语和术语之间语义关系的集合,是支持本系统知识项之间基于语义自动关联的核心组件。例如在数据库领域,“数据库”、“锁”、“数据库管理系统”、“DBMS”等都是术语,术语之间存在多种语义关系,如同义关系、对义关系、反义关系、同位关系、上下位关系、部分整体关系等。在本体中,“数据库管理系统”与“DBMS”就是同义关系。而概念是以文字形式的术语所描述的超出文字的意义,同义的术语表示相同的概念。

(2)知识点、知识的形态与知识项

知识点是教学中常用的一个概念,它是从教育和学习的角度对一个领域知识进行标志和处理的基本单位。一般在教材编写中将一个小节作为一个知识点。在概念上,本系统也以知识点作为知识处理的单位之一。

知识还有不同的形态，文字、图形、表格、视频等是知识表示的不同形态。本系统将知识的不同形态分别进行存储和处理。

知识项是知识点的具体化，是知识点与知识形态的结合。本系统中知识点和知识形态都是抽象的概念，对于给定的知识点只有通过具体的形态表示出来才被具体化。知识项是本系统对知识处理的最小单位。

(3)设计理念

将领域知识分为3个层次：最高层用本体实现领域概念知识的横向(关联)和纵向(层次化)语义关系网，在第二层用知识点描述领域概念知识的有效组合，第三层通过知识项表示知识的具体形态。这三层实现了由整体到局部、由抽象到具体领域知识的语义关联和层次化框架。

在语义层面，应用本体作为领域知识的概念骨架，通过本体建立系统中各个知识项之间的语义联系，基于此语义联系帮助读者/研究者动态检索相关的知识项。在数据的存储层面，将整体的文档知识按知识点和知识形态存放在数据库中。在实现层，将自动化技术和人工处理相结合，应用自然语言处理技术自动地识别和提取领域专业词汇，并应用互信息等技术识别领域词汇之间的聚类关系，再辅以人工的鉴别和修正，达到效率和质量的折衷。

1.2 系统基本功能

为让读者对系统有一个整体了解，先从外观上介绍系统的功能。系统可视为一个服务系统，在功能上可分为服务准备和服务实现2个部分。本文以数据库领域为例介绍本系统功能。

服务准备部分的主要功能是根据不同主题或领域建立服务项目，针对每个服务项目选择和导入相关的电子文档(知识资源)，根据所选领域中导入的知识资源自动学习和创建领域本体，包括领域词汇和词汇之间关系的确认。在对领域词汇关系的精确度要求不高的情况下，这一部分工作可由普通系统维护和管理工作人员进行，当要求精确地确定词汇间关系时，需要领域的专业人士进行人工调整和修正。

图1是对所导入数据库领域的电子文档进行自动识别领域词汇后，由领域专家(或教师)对它们之间的语义关系作进一步确定和编辑的用户界面。

服务实现部分由学习者操作，学习者选取要学习的领域，以该领域某个知识源为主线查看其中的知识点，而可用相关资源的链接将自动出现在导航窗口中，实现同时阅读和参考多个知识资源。在图2中，学习者选取了《数据库基础及应用》这一本书作为学习主线，在学习‘数据模型’的主要数据模型这一

图1 服务端的专业词汇关系构造器用户界面

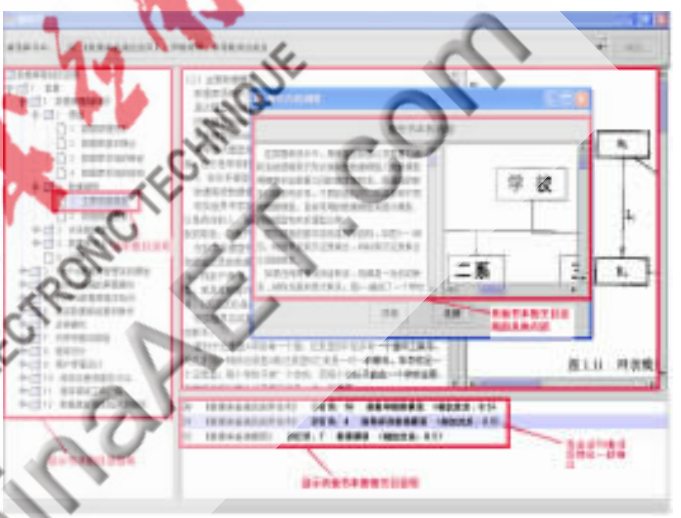

图2 客户端智能辅助学习系统界面之一

小节时，相关的内容在主窗口显示(图形和文字分开)，在主窗口下方的导航栏显示了在其他书本和章节中相关内容的链接。任意选取其中一个链接，将显示其中的详细内容。

2 基于语料库自动学习领域本体

本设计应用基于语料库的自然语言处理技术从电子文档资源中识别领域专业词汇，用互信息技术分析领域词汇之间的可能关系，再辅以人工鉴别和修正。图3说明了用计算机辅助本体构造过程的3个基本步骤：选

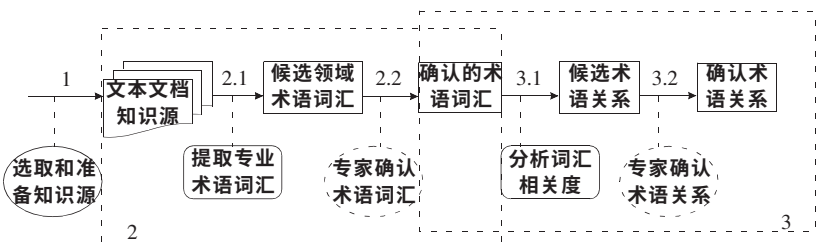

图3 本体学习流程

择和准备知识资源、提取领域术语词汇、建立词汇术语之间语义关系。图中以圆角矩形表示的步骤由系统自动完成；用椭圆表示的步骤由系统提供人机交互界面，以人工操作完成；虚线的人机操作在精度要求不高时可以省略。

2.1 知识资源预处理

领域本体学习的第一步是准备电子知识资源，电子版的教程是较理想的。这一步的基本任务是导入电子版的知识资源，并对导入的资源格式进行规范化处理，转换为系统可识别和处理的知识项单元。文字部分被转换为 text 格式，图形部分统一为通用的 jpg/bpm 等格式并加上图形的标题存入数据库中。这一步不需要深入的领域专业知识，可由一般系统服务人员进行。

2.2 识别领域词汇

第二步主要实现从所导入的文本知识资源中识别领域词汇，并最终确认。这一步主要是取得领域词汇的词干，即构成领域术语的最基础元素。一般文本挖掘方法在识别词汇时事先筛选某些常用词汇作为高频词，它们在识别过程中被排除。这里不采用此方法，因为中文的领域词汇通常也会使用某些常用字/词，对它们赋予新的领域含义。本系统应用基于语料库[2]的自然语言技术，先对文本的知识资源进行中文分词处理[3]，再对所出现的词汇进行词频分析，将资源文档中的词频与语料库词频进行对比，频率显著高于语料库中的频率时，推断它为领域词汇。由自动系统识别出领域词汇后，可由领域专家再进行确认和修正。

2.3 识别词汇关系及组合术语

第三步的目的是识别并确认领域词汇之间的关系，根据它们之间的有效组合并得到领域术语集合。

从构词法上看，专业领域中的词汇有 3 种基本构成形式：给普通词汇赋予新的领域含义；创建一个全新的词；以前两种形式为词干加上前缀或后缀形成新词。

第一种领域词汇通过分词系统自动划分为一个独立的词，在语料库中也会出现，它可通过上一步的词频对比分析识别得到。第二种领域词汇在自动分词系统中无法分出，在语料库中也没有该词。它由若干单字或常用词组合而成。第二种和第三种可应用信息论中的互信息，自动地从样本文档中识别。信息论中互信息反映了一种信息与另一种信息相关联的程度，用下式表示：

$$M(a,b)=\log 2(P(a|b)/P(a))$$

其中 $P(a)$、$P(b)$ 分别表示事件 a 和 b 出现的概率，$P(a|b)$ 为事件 a 相对于事件 b 的条件概率。在本系统中，以样本文档中总词数 cntTotal 为基数，以词出现的次数 c 除以总词数作为概率估计值。$P(a|b)$ 用 a 与 b 同现次数除以 b 出现次数作为估计值。仅对文档中先后同现 2 次以上的词进行互信息统计分析，应用互信息计算公式通过编程计算得到词汇两两组合的相关度表。以词汇之间的组合关系为边，以相关度为权值构造一个有向加权多图。图 4 就是对数据库电子文档应用互信息计算得到的加权图之一。根据它就可以在一定置信度范围内获得词汇之间的可能组合关系。

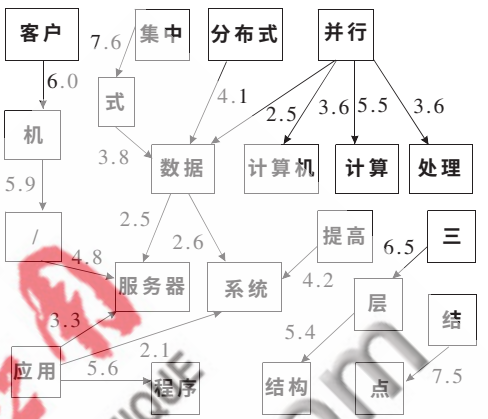

图 4 词汇高频组合关系图

词汇的组合关系蕴含着语义关系。基本的语义关系包括同义、上下位、反义、对义、部分与整体关系等。对这些关系还分别赋以一个相关度值，以反映它们之间关联程度。自动建立了所识别词汇之间的组合关系后，赋予词汇之间默认的关系和相关度值。有领域经验的人可对这些关系和相关度值进行编辑，在实际的辅助学习平台应用中由教师进行操作。图 1 就是实现此功能的操作界面。

本文介绍了一个应用所构造本体的智能辅助学习系统的功能、设计想念和实现方法。通过该系统它可将一个学科(领域)的多种资源存入数据库中，实现学习某一主题的知识时，可以同时对比阅读多种相同或相关主题的内容，省去手工查阅多种资料的麻烦，还可直接跳转到另一种资源，以它为主继续学习，这给学习和研究带来很大方便。下一步的工作是将该技术应用到企业知识管理中。

参考文献

[1] 周宁，张玉峰，张李义，等.信息可视化与知识检索[M].北京:科学出版社，2005.

[2] 北京大学计算语言学研究所.人民日报语料库[DB/OL].[2001-05-10].http://www.icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp.

[3] 张华平，刘群.计算所汉语词法分析系统 ICTCLAS，http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/，2002.

(收稿日期：2010-02-02)

作者简介：

李向阳，男，1971 年生，副教授，主要研究方向：知识处理与知识工程、语义 Web。