文章编号:1006—9860(2015)03—0081—09

基于本体学习算法的学科本体辅助构建研究*

——以学习元平台语文学科知识本体的构建为例

丁国柱，余胜泉

(北京师范大学 教育学部 教育技术学院，北京 100875)

摘要：人工构建基础学科领域本体是一项复杂的工作，结合领域专家和本体学习是较快建立学科本体的方法。该研究在分析基础教育语文学科知识分类体系基础上，把语文学科知识本体分为语文知识本体、语文教材本体和语文学科教学本体，构建了语文学科的骨架本体，同时利用TF-IDF算法实现本体学习，提取语文学科中的关键概念和关系，完善语文骨架知识。该研究依托学习元平台实现，从一定规模的领域文档中获取关键概念及其关系，成功构建了包含1303条概念、136条关系、269012条实例的语文学科本体库，并开展了语义检索和学习资源标注应用。

关键词：本体；本体学习；语文学科本体；语文学科知识体系

中图分类号：G434 **文献标识码：**A

一、导言

基础教育学科知识体系是在课程标准指导下，学科知识点之间相互关联构成的知识系统，是学习资源深度聚合、情境导航、资源推荐、智能导学的基础，是基础教育在线学习平台的核心技术和难点。当前不少基础教育在线平台都推出了自己的知识点体系，然而这些知识体系描述相对粗略，缺乏教学信息的描述，知识点之间只是单一维度的上下位线性关系，缺乏知识之间多维度逻辑关系的体现。在数据系统中表征相对完整的学科知识体系，其实质是构建学科知识本体。本体是某个领域的概念集合和概念之间的关系，通过形式化的方式表达该领域的知识及其知识结构。构建学科领域本体是一个浩大的工程，单纯靠领域专家手工构建，人力和时间成本都不小，通过计算机自动构建本体是解决这个问题的关键，国外把这种技术称为本体学习(Ontology Learning)。然而，单纯靠机器自动构建的准确率还较低，当前也没有一个成熟的本体完全通过机器自动构建完成，在中小学学科领域也还尚未形成可参考的本体构建思路和模型。此外相比起自然学科注重概念及学科逻辑，一些人文社会学科的实例数量相当庞大，实例的填充和完善也是本体构建的重要工作。

相比起其他知识本体，学科知识有其独特的性质，不单要考虑知识内容本身，也要考虑学科教学内容，教学内容组织形式的不同决定着教学开展顺序的不同。同时，中小学领域不同年龄段学生认知水平和能力的差异较大，教学也要依据学习者的不同层次展开。此外，教学的开展也依赖于课程标准和学科教学规律。因此，本研究从语文知识、语文教材组织、语文教学(包括教学目标和学习者)三个维度出发，通过分析学科知识分类体系，导入语文教材体系，结合课标要求和语文学科教学法，由领域专家构建的骨架本体，结合本体学习从领域文档中提取新术语，并利用关联规则在相应的术语和骨架本体中的概念之间建立层次关系，导入概念实例，建立知识内容、知识组织和学科教学关联的学科知识本体，实现较快速度地构建学科知识本体，并以学习元平台[1]为例，展示中小学语文学科知识本体构建流程。

* 本文系“移动学习”教育部—中移动联合实验室建设项目(项目编号:移有限技合同[2012]934)研究成果。

二、相关研究

(一)学科本体构建

20世纪90年代国外学者就对学科本体的构建展开研究，用形式化本体描述学科[2]，具体的学科领域涉及生物学科[3]、化学学科[4]等，此外还有K-12本体[5]、学习者本体[6]等。我国学者也对学科本体构建展开了研究，如佘莉构建了几何学科知识本体[7]，詹川研究了教育心理学的本体构建[8]，郝兴伟构建了计算机基础学科本体[9]，魏顺平构建了教育技术学学科本体[10]，顾芳从领域分析、本体和知识的表示、本体和知识的分析三个方面提出MDOS方法构建一个多学科领域本体[11]，杜小勇研究了学科本体的构建方法[12]，提出本体需要进化以适应领域知识的增加，赵呈领等基于初中物理课程构建了物理学科的领域本体库[13]。在中小学领域，学科知识本体构建有初步的进展，如王冰洁构建了小学英语学科本体[14]、郭军梅构建初中语文本体[15]等。

从上述研究可以看出，研究者不约而同地从教学大纲或课程标准出发建构学科本体，但没有考虑教材组织的不同对学科知识本体构建的影响，同时缺乏对学科教学本体的研究。此外，这些研究主要通过领域专家手工构建本体，构建成本很大。另外一些研究基于Protégé构建本体，如果本体容量超出JDK内存限制就会出现内存溢出导致无法使用。在实际应用中含有大量实例的人文社会学科本体无法在Protégé上运行，构建的本体要在互联网领域内应用还需要作出很大努力。

(二)本体学习

本体学习是指从数据源中提取概念及其关系，数据源主要包括计算机可以识别的各种文本数据，如HTML、XML、OFFICE文档、数据库文件等。数据库是规定好格式的XML文档，属于结构化、半结构化数据，可以通过模式识别或者基于模板的方式获取概念；而对于占较大比例的非结构化数据，目前主要有3种概念获取方法。

1.基于语言学的方法。利用语言规则编写特殊词法结构或模板来提取相关概念，这个方法依赖于具体的语言环境，适合提取概念与概念的层级关系，李丽双的博士论文中有较详细的阐述[16]，该研究以同义关系、并列关系、上下位关系、相反关系为例，对应的语言规则如下：

(1)同义关系：描述一个概念的同义关系时常用到“又称/又称为/简称/又叫做/叫做/亦称/亦称为/别称有”等关联词，当句子中出现这些关联词时，被描述的词汇很可能就是和主语等价的概念。如“词又称曲子词、乐府、乐章、长短句、诗余、琴趣”，根据上述规则，计算机可以发现词的等价类有曲子词、乐府等。

(2)并列关系：描述一个概念的并列关系时常用到“和/或者/或/与/以及/还有”等关联词汇，如“人物外貌描述和人物心理描写是刻画人物的方式”，人物外貌描述和人物心理描写是并列关系。

(3)上下位关系：描述一个概念包含的子概念时，常用到“包含/一共有/主要有/可分为/按(种类、性质、等级、时代、风格)分有”等关联词汇，如“古诗可分为古体诗和近体诗”，那么古诗就下位概念包括了古体诗和近体诗。

(4)相反关系：描述2个相反概念时，常用到“相反的是/与之相反的是/与之不同的是”等关联词汇，如“与褒义相反的是贬义”，可以提取褒义与贬义的相反关系。

2.基于统计的方法。这种方法主要根据领域概念与普通词汇拥有不同的统计特征，如关键概念会反复在领域文档中出现，相关度高的概念会在多个文档的同一句子中出现等，常用的算法有互信息、信息熵、TF-IDF算法等。

3.混合方法。结合语言学和统计学方法，单纯的基于语言学方法或基于统计学的方法在召回率上可能不太令人满意，结合两种方法的优势可以提高概念的提取效果。

无论哪种方法来实现本体学习，相当数量的领域文档是必不可少的，恰当的领域文档有助于本体的自动构建。

(三)关联规则

用于发现不同集合间的关联程度，是数据挖掘重要的课题。当新概念被提取并与其他概念共同出现在一些领域中时，可以认为它们之间可能存在关联。当关联程度满足最小支持度阈值和最小置信度阈值时，新概念与骨架本体中的概念就可以建立关联。

三、中小学语文学科知识本体构建思路框架

本体学习能够利用机器从海量领域文档中提取相对数量的概念，然而到目前为止还无法通过机器构建一个足以和领域专家相媲美的本体，经机器处理生成的本体往往都要人工干预后方能使用，所以由领域专家构建骨架本体，机器辅助完善本体是一个较为可行的方案，整体框架如下页图1所示。

本研究中以中小学语文学科本体构建为例，整体研究路线描述如下：

(一)选定学科，准备学科材料

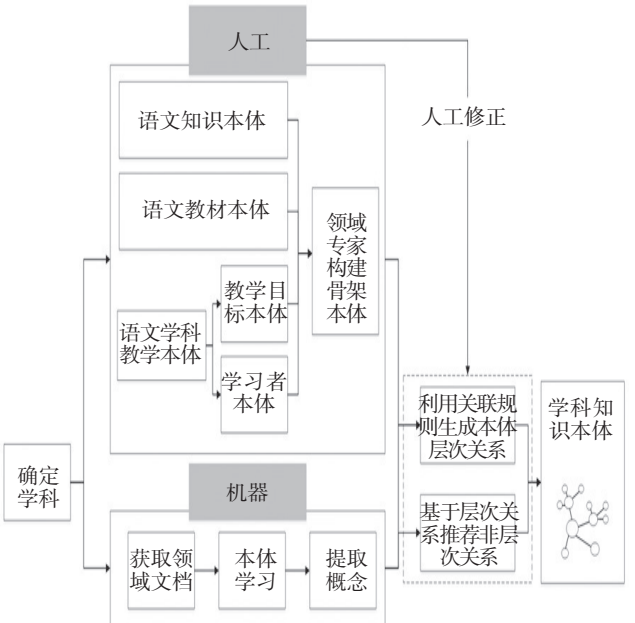

图1 学科知识本体构建思路框架

本文以语文学科为例。不同的学科需要不同的领域专家参与，同时确定了领域文档的范围：教材、教辅、教研文章、试卷练习、教学设计、教案、学情分析、教学笔记、教学论文、学习心得、教学课件等是领域文档的重要来源，为了能更全面地挖掘到学科概念，还可以将领域文档按照不同学段分开，提高不同学段中重要概念的提取覆盖率。

(二)领域专家构建骨架本体

分析语文学科知识分类体系，得到语文学科主要知识分类及其组织形式，如语文学科主要的知识有字、词、句、篇、章、修辞、语法等基本概念，每个基本概念又包含若干重要概念，如字包含读音、偏旁、结构，句子有表达方式、句子结构、修辞手法等子概念。导入不同版本的中小学语文教材，深入分析语文课程标准，提取课程标准中的重要知识点，结合语文学科教学法，构建语文学科知识骨架本体。

(三)利用本体学习完善本体

1.获取学科领域文档，领域文档可以编写垂直爬虫，从互联网上抓取。经互联网获取的数据量较大，对获取的文本进行清洗，用正则表达式去除HTML标签，得到纯文本文件。

2.利用本体学习算法从领域文档中提取概念，本研究采用TF-IDF算法实现[17]。TF-IDF算法是一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降，其中TF(Term Frequency)表示词频，对于k个文档中某一个特定词汇t_i，在其重要程度可以表示为：

$$tf_{i,j}=\frac{n_{i,j}}{\sum_n n_{kj}} \quad (1)$$

其中，$n_{i,j}$表示该词在k个文档中出现的总次数，$\sum_n n_{kj}$则表示k个文档所有的词汇总数。

IDF(Inverse Document Frequency)表示逆文档率，用于总文档与计算出现词汇的文档总数的比例，再取对数得到，用公式表示如下：

$$\mathrm{idf_i}=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \quad (2)$$

对于这个特定词汇t_i的TF-DIF值为$\mathrm{tfidf_{i,j}}=\mathrm{tf_{i,j}}\times\mathrm{idf_i}$。

3.利用语言学方法，结合骨架本体构建概念层次关系。以同义关系为例，将领域文档以逗号和句号进行分割，获得有同义关联词汇(又称/又称为/简称/又叫做/叫做/亦称为)的句子，其格式为：Concept1又称/又称为/简称/又叫做/叫做/亦称/亦称为 Concept2，系统建立这两个概念的等价关系。

(四)关联规则

关联规则可以获取没有在领域文档显性表达的概念之间关系，本研究基于Kinshuk对于关键词关联紧密度的算法[18]：

1.在文章中列出的任何一个关键词都代表该文章的一个重要概念；

2.如果两个关键词出现在同一篇文章中，那么它们之间必然存在某种关系；

3.如果两个关键词出现在同一篇文章的频率越高，那么它们的关联性更强；

4.在同一个句子中，两个关键词的相距越近，那么它们之间的关联越强。

Kinshuk的算法只用于揭示关键词之间的关联程度，并未能发现它们之间的层次关系。Sanderson等人假设对于任2个词w1和w2，如果w1总是出现在w2的上下文中，即w1和w2一起出现而且w1出现的概率要高于w2，则可以把w1是w2的统计上位词[19]。因此在足够全面的领域文档中上位词出现的几率要高于下位词[20]，如通过关联规则算法可以发现“人物描写”与“概念外貌描写”“语言描写”“行动描写”“心理描写”关联程度较高，但“人物描写”出现的频度要比其他四项高，可以假设“人物描写”是其他四个概念的上位概念。

四、语文学科骨架本体构建

如前文所言，语文学科知识本体包含语文知识

本体、语文教材组织本体、语文学科教学本体(包括教学目标本体和学习者本体)，如图2所示，本研究从这三部分进行语文骨架本体的构建。

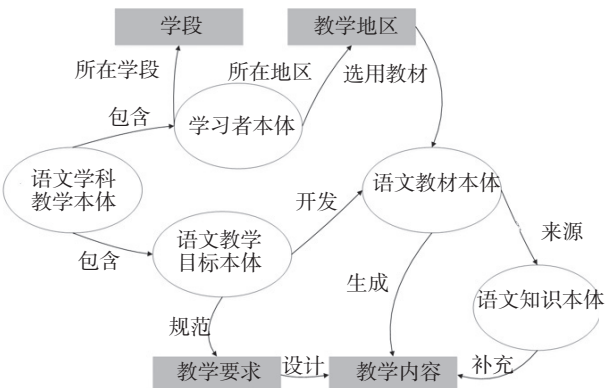

图2 语文学科知识本体模型

(一)语文知识本体

知识分类是构建知识本体的前提，目前比较有影响的语文学科知识分类体系如下。

1.魏书生的语文知识树[21]。20世纪70年代他将语文知识画成结构图，主要以树型结构来显示，将语文知识分成基础知识、阅读与写作、文言知识、文学常识四个大枝干，在此基础上再细分成22个子分支。这个知识体系对我国语文教育有较为深远的影响，使得语文学科脉络变得清晰可见，在教学过程中教师利用语文知识树可以方便地对知识进行梳理。然而，这种划分方法建立在当时的语文教材体系上，如果语文教材不合理，那么这个分类体系也就不合理了。此外，这种分类更多体现语文中的显性知识，对蕴含在“冰山下面”大量隐性知识的分析还显不足。

2.韩雪屏认为语文的课程教学目标体现了语文课程的价值观和立场[22]，决定了语文的知识内容和类型。结合现代认知心理学的广义知识观，他认为语文知识分类包含了陈述性知识、程序性知识和策略性知识。这种知识分类体系有利于知识的扩展和延伸，正如他说的“从一个个孤立的文篇，向大文化概念扩充……将促进语文课程整体知识类型的转化”，但其知识分类在本体构建时存在一些鸿沟，一般用户需要先掌握陈述性知识、程序性知识和策略性知识背后的内涵，方能进一步完善本体。此外，一些隐性知识不容易归类，这种分类法来构建本体有点困难。

3.王荣生对语文知识体系有长期持续地研究，在其博士论文《语文科课程论建构》中，突出了知识技能“听、说、读、写”，也提出语文中“潜层面”的知识——缄默知识[23]，这个名词和隐性知识是异曲同工的。他后来还提出语言学、文学、文章学和心理学，是支撑语文学科的四根支柱[24]，由此构建语文学科体系。

4.李海林从知识类型概念出发，提出语文知识可以从多个角度观察：一是以认知心理学为基础的陈述性知识、程序性知识和策略性知识；二是以加涅知识体系理论为基础的现象知识、概念知识和原理知识；三是无意识知识、言述性知识[25]。前两种方法的分类有较坚实的理论基础，严谨性比较强，第三种分类方法有一定的见解和创新性，深入地理解，也可以看成隐性知识的一种。

5.邱福明从认识论基础和存在论的视野出发，认为语文知识体系应该从内容维度、形式维度和意义维度进行分析，在其博士论文《语文课程知识的存在论研究》中提出根据人性论观点和整合意义视角，尝试构建基于促进意义实现的语文课程知识系统[26]。该研究在总结前人的基础上，对潜藏在语文知识体系中的隐性知识的挖掘有了更进一步的研究。

综合多位语文教学专家的研究，可以得出语文学科知识体系不但包含单纯的工具性知识，还有人文性知识、价值性知识，是工具性和人文性的统一。对于构建语文学科知识本体的启示是：既要考虑工具性的“字、词、句、篇、语、修、逻、文”，也要考虑语文本身蕴含的人文性，如中国源远流长的文化、宗教、风俗、艺术、哲学里包含的情感性知识、态度性知识、价值取向的知识、思想性知识等，应该从多个综合维度考虑。本研究主要从语文的工具性知识和人文性知识两个维度构建骨架本体。

1.工具性知识可以参考王冰洁研究的小学英语本体的构建方法，从课标中提取最关键的概念，包括汉字、词汇、句子、修辞、逻辑、语法、文章、文学等。考虑到实例填充时部分概念可以被其他概念使用，其实例也依赖于其他概念，如修辞和语法的实例依赖于句子，文学依赖于文章。如果将这类关键概念单独作为一个维度来构建本体会产生大量的数据冗余，同时也增加了后期对本体维护的成本，因此可以对一些关键概念进行裁剪，最后确定汉字、词汇、句子、文章、作为工具性知识的顶级概念，由这些顶级概念的实例关联其他关键概念如修辞、语法等。

2.人文性知识到目前为止还没有一个较为权威的分类，薛为春认为人文知识是人类认识、改造自身和社会的经验总结，人文精神则是人文知识化育而成的内在于主体的精神成果[27]。人文的核心是重视人的文化，相比工具性知识，人文性知识具有一定的主观性，不同人看待一件事物的角度和态度有

可能不同。人文性知识更多反映在教学三维目标中的情感态度价值观。

(1)情感与态度：表征语文教材中人们对于事件、事物、人物等内容的情感和态度，如人教版《桂林山水》的“桂林山水甲天下”蕴含的人文性知识是对祖国壮丽河山的热爱；再如语文版高二课文《论语·己所不欲，勿施于人》反映的就是为人处世的态度。

(2)思想和价值观：表征语文教材中人们对于事件、事物、人物等内容的思想及其反映出来的价值观，如人教版《过零丁洋》中文天祥名句“人生自古谁无死，留取丹心照汗青”反映的是爱国思想；再如陶渊明《归园田居》，反映的是作者诗人淡泊宁静、志趣高洁的思想和价值取向。本研究认为人文性知识可以通过人、事、物、情表达出来，如松、竹、梅被古人誉为“岁寒三友”，不少文人墨客留下名句借喻它们来表达高尚的情操，而表达自己情感的背后可能有一些相关事件，如此人、事、物、情就有机地联系起来。

(二)教材组织本体

中小学语文教学中，语文知识的组织形式通过教材呈现，一线教师和学生最终通过教材完成对语文知识的传授和学习。如果说语文知识本身是内容，那么教材则是对内容的组织，如李白的《静夜思》，在唐诗三百首中是一种组织形式，在不同语文教材版本中又是另一种组织形式。把语文知识内容与教材课文关联有助于学习资源的组织，学生学习往往以教材教学顺序展开，学科知识本体就可通过教学流程进行整合，为学习资源深度聚合、情境导航、资源推荐、智能导学等应用打下基础。同时，也较好地解决同一个语文知识出现在不同教材版本的语文教材中的情况，使得不同教材都能用到同一个语文知识相关的资源。语文教材组织本体如图3所示。

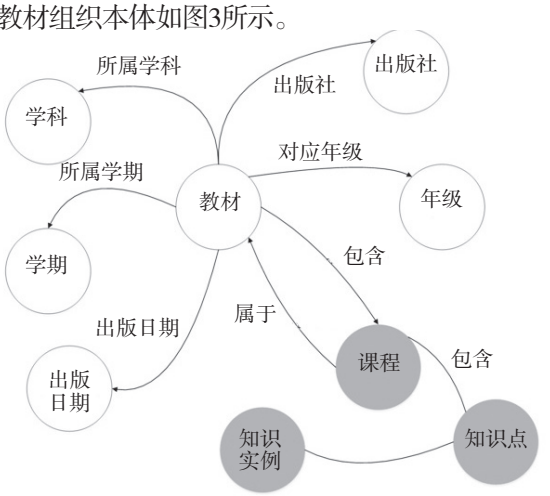

图3 语文教材组织本体

(三)语文学科教学本体

只有语文知识内容本体和语文知识组织本体还不能反应语文知识应该如何教授，也无法表征不同层次的学生对同一个知识点的教学要求，因此必须深入语文课程标准，认真分析其对语文知识的选择、对不同学习层次学生的要求、教学方法的指导意见等，这就是语文学科教学本体。同一个知识点，小学二年级的要求是记忆，高中二年级则可能是评价；同一个知识点对不同年龄阶段学生适合采用的教学方法也可能不一样。语文学科教学本体包含相关知识实例的教学方法、学习方法、教学目标、教学内容、教学测量和评价等。教育部在2011年颁布《义务教育语文课程标准》和2012年颁布《普通高中语文课程标准<实验>》，共同组成了中小语文的课程标准。从课程标准的行文逻辑可以看到课程性质、课程基本理念、课程设计思路是整个语文学科课程设计、教学、评价、资源制作的起点，指导着课程目标与内容的构建。

不同学段的教学目标与内容规范了语文知识内容的选取和语文知识组织的形式。课程标准根据不同学段学生的认知水平和学习能力，针对不同学段的学生设定了学段教学内容与目标，其中义务教育阶段为“识字与写字”“阅读”“写作”“口语交际”“综合性学习”；普通高中阶段为表达与交流、诗歌与散文、小说与戏剧、新闻与传记、语言文字应用、文化论著选读与专题研讨、阅读与鉴赏、高中语文基础知识等，各个学段相互联系，螺旋上升，最终全面达成总目标[28]。学习者所在区域和使用的相关教材也可以关联起来。语文学科教学本体模型如下页图4所示。

(四)语文学科骨架本体结构

综上所述，课标规范了语文知识内容和针对不同学段学生的教学目标与要求，教材根据课标的指导，将语文知识内容按照一定的方式组合排列形成相应的知识内容组织本体。一个较完整的语文本体，应该对每一个知识实例说明其知识类型，在什么教材中出现，对不同学习能力、不同年龄层次的学生的学习要求及教学方法，基本能够指导老师如何教学，指导学生如何学习。学科知识本体的骨架系统如下页图5所示。通过课程标准的教学教材内容规范教材编写，而教材内容的来源则是语文的学科知识。同时，学段教材内容、教学目标和实施建议指导语文教学、教材编写、课程资源开发等工作的开展。

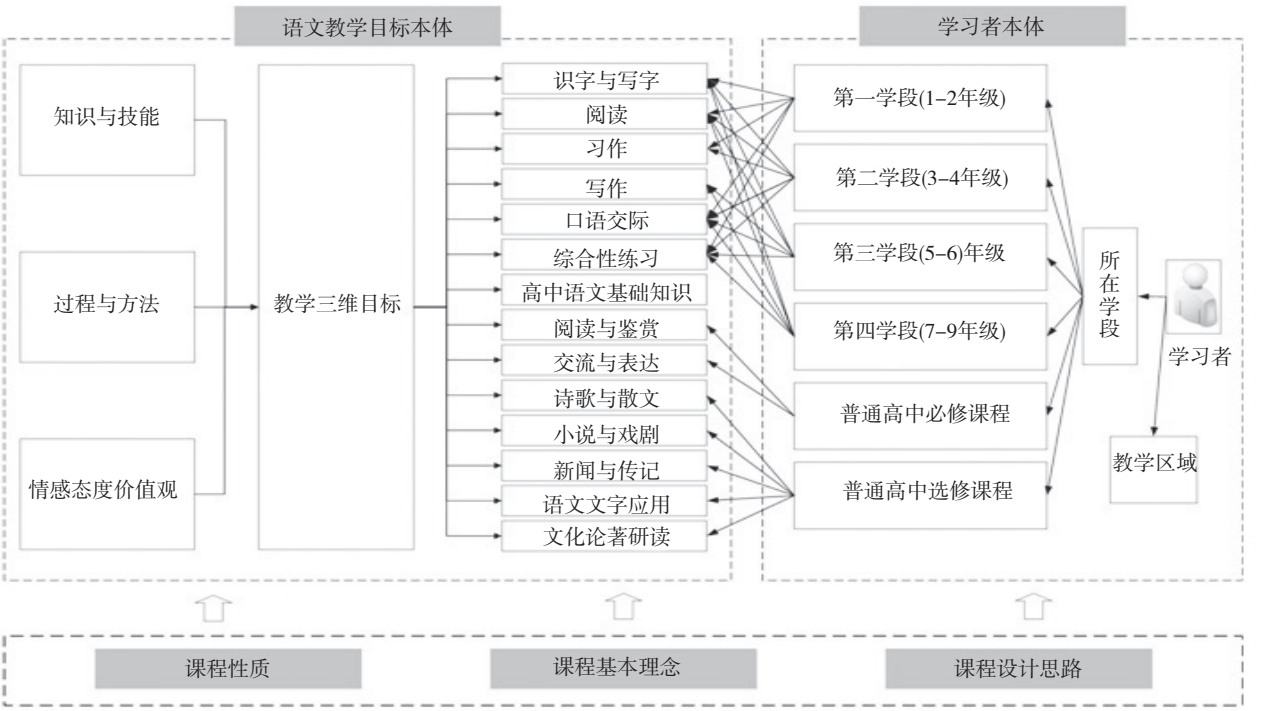

图4 语文学科教学本体

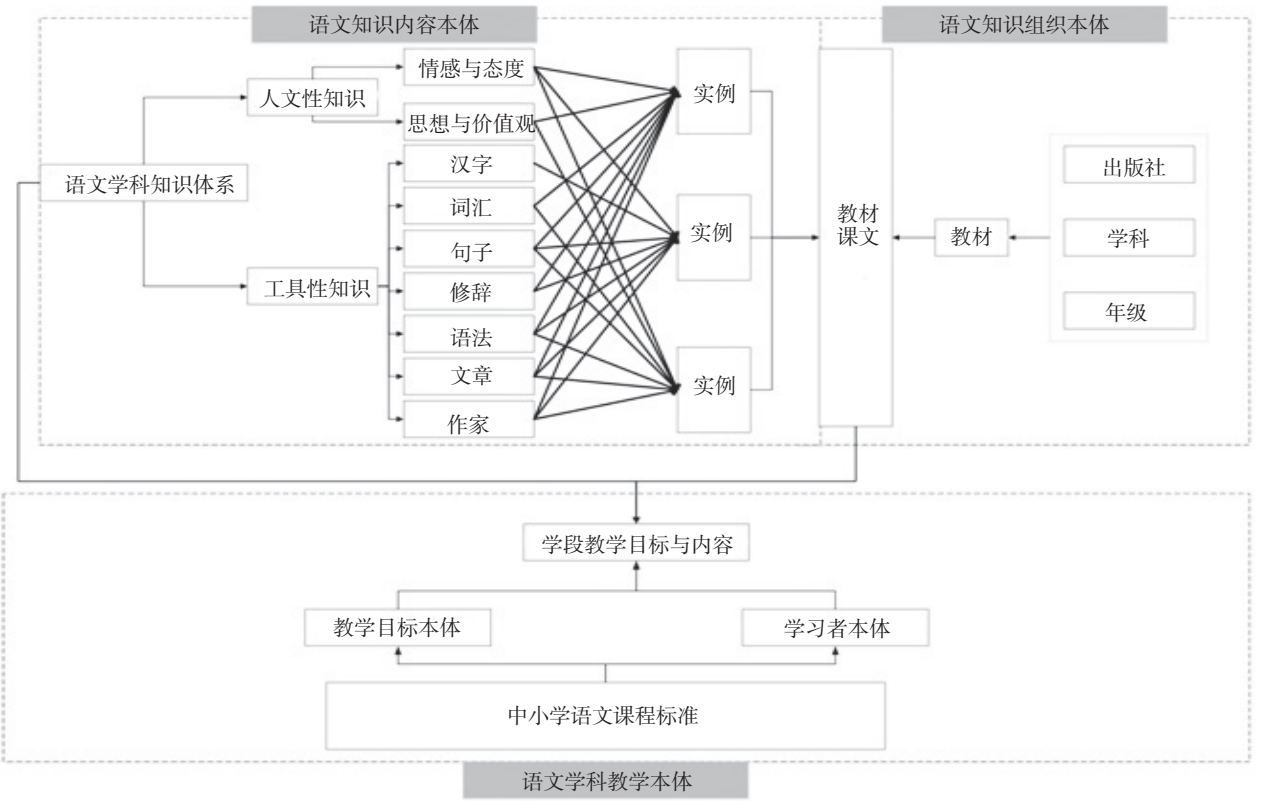

图5 学科本体骨架系统

五、语文学科知识本体的辅助构建

本体学习可以利用相关算法实现本体的自动构建，能够辅助构建语文学科知识本体，降低人力成本和提高本体构建的效率，本研究基于学习元平台

进行本体学习实现语文学科知识的辅助构建。利用Protégé构建好语文学科骨架本体、语文知识内容本体，包含汉字、词汇、句子、文章、作家等5个维度，由这5个维度关联其他264个核心概念；根据出版社、学段(年级)、学科组成导入教材信息，每本教材细化到每一个课文(即教学内容)，包括人民教育出版社、北京师范大学出版社等246本语文各年级教材，共9271门课程；根据语文课程标准，提取主要的教学内容和目标，建立以课标为中心的知识点体系，分小学语文、初中语文、高中语文三个领域，共605个核心概念。每一个维度建立一个本体文件，通过JENA导入至语义数据库JEAN TDB，同时为了方便Web环境的展示，支持本体可视化，将本体用树状形式显示。经过计算机自动提取概念后，经过初步的手工筛选，主要概念由874个增加到1303个，关系共136个。本研究本体学习算法实现流程如下。

(一)领域文档获取

本体学习的基础是相对全面和数量较大的领域文档。本研究一方面基于学习平台中的语文学科资源，另一方面编写垂直爬虫，从专题语文学习网站、教研网站、语文知识网站、成语网站、在线字典、在线古诗词等网站中获取资源，半结构化的数据如成语网站、字典网站存入数据库中；非结构内容如教学论文、教学设计、试题、课文等，抓取数据后去除HTML、CSS、JS后以纯文本方式储存。爬虫抓取的内容可能存在重复，在储存文本时，检测文本内容的重复性，减少重复的文档对本体学习的影响。为了提高本体构建质量，文档要求尽量结合实际教学，同时能覆盖主要的语文知识。

(二)利用TF-IDF算法提取概念

TF-IDF算法前文已描述过，其实现流程如图6所示。首先读取爬虫获取领域文档，概念一般为名称、动词、形容词等实词。为提高计算效率，程序利用分词工具把副词、介词、连词、助词、语气词等虚词去除，只留下实词，本研究用到的分词工具为ANSJ，基于中科院的 ICTCLAS 中文分词算法的开源Java分词工具[29]。利用TF-IDF算法从4000篇领域文档中初步得到2231个关键词，包括字形、拼写、同义词、反义词等关键术语。然而，没有优化的TF-IDF算法提取的关键词多是单个词汇形式的概念，两个词组成的复合术语提取效率不佳，如修辞手法、象征性意象、明用历史故事等。分析发现组合性的术语一般是成对出现，而且存在先后顺序，如修辞和手法出现时，程序对“修辞”提取的位置为N，那么“手法”提取的位置即为N+1，这2个词共现的频率较高，程序可以将其设置为可推荐的关键词，一般组合概念由2到3个词共同组成。因此，程序在提取关键词时，也记录该关键词出现的位置，如词汇X1，X2，X3同时出现，而且X3(位置)-X2(位置)-X1(位置)=1，那么X1X2，X2X3，X1X2X3都可能是一个术语。经过优化后，增加了独体字结构、人物描写记叙文、汉语拼音拼写规则、山水田园诗人、第一人称、形容词短语等36个新概念。

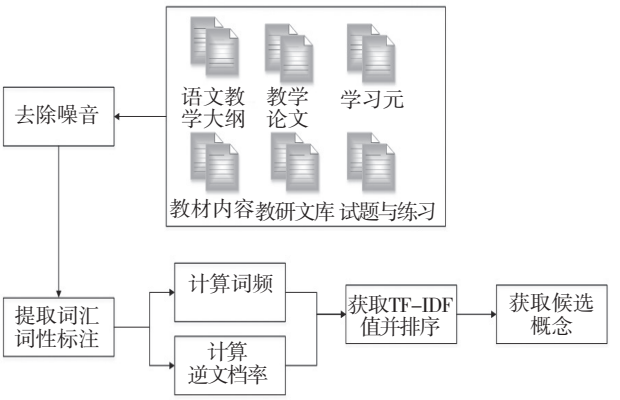

图6 利用TF-IDF算法提取概念流程

(三)本体概念关系提取

概念关系包括层次关系和非层次关系两种，层次关系利用汉语语言规则和关联规则提取，而非层次关系则基于提取好的层次关系，向本体工程师推荐生成。

1.利用汉语语言规则概念层次关系

首先编辑好反映等价关系的词汇模板，如等价关系的“Team1又叫做Team2”；然后对领域文档根据句号进行切分，匹配等价关系关联词“又叫做”，提取关联词2侧的词汇，标注为概念，并建立等价关系。并列、上下位、相反关系的提取也按照这种算法思路提取。算法流程描述如表1所示。

表1 基于汉语语言规则的概念层次关系提取算法

输入：领域文档 输出：概念及其层次关系三元组<Concept1,Relation,Concept2>
关键步骤： Step1 读取(等价/并列/上下位/相反)词汇模板 Step2 读取领域文档 Step3 把文档以句号为标志分开成句子 Step4 根据词汇模板，提取包含相关关系词汇的句子 Step5 根据TF/IDF算法提取的概念以及词汇模板定义的关系，生成概念及其层次关系三元组 Step6 检测该三元组<Concept1,Relation,Concept2>在本体中是否已经存在，如果不存在则加入本体

2.一些上下位关系的概念无法通过词汇模板提取，可以利用关联规则提取。

前文TF-IDF算法提取概念，记录概念频率，根据关联规则提取符合规则的词汇(本研究中把最小支持度设为0.5，最小置信度设为0.6)，把频度较

高的词汇置为上位词，频度较低置为下位词。算法流程描述如表2所示。

表2 基于汉语语言规则的概念层次关系提取算法

输入：领域文档
输出：概念及其上下位关系三元组<Concept1,上位关系/下位关系,Concept2>

关键步骤：
Step1 获取概念集合C={C_1,C_2,C_3…C_n}，并依次读取概念C_i
Step2 获取与概念C_i在文档集合D中共现的其他概念C_x($C_x \in C, C_x \neq C_i$)
Step3 判断C_i和C_x是否已经存在关系，如果存在就跳出，返回Step1计算C_{i+1}其他概括的关系，如过不存在，进入Step4
Step4 计算只包含C_i或C_x、或者C_i和C_x一起出现的文档数：Count
Step5 计算C_i与C_x关联的支持度：$\partial=\frac{C_i \wedge C_x}{Count}$
Step6 比较C_i与C_x在包含C_i、C_x的文档数集合中出现频率的大小，把频率较大的数字为Confidence
Step7 取频率较大的假设为上位概念设为C_s，频率较小的为下位概念C_a
Step8 计算Ci与Cx关联的置信度：$\theta=\frac{C_i \wedge C_x}{Confidence}$
Step9 判断C_i与C_x关联的支持度和置信度值是否达到系统预设值，如果达到则输出<C_s，下位概念，C_a>和<C_a，上位概念，C_s>

实现上述算法后，系统提取了等价、并列、上下位、相反关系共98个，如表3所示。

表3 层次关系提取结果

等价关系	词—长短句—曲子词—曲子—曲词—乐章—琴趣—诗余，衬托—映衬，顶真—顶针—联珠，绝句—截句—断句，押韵—压韵，通假—通借等
并列关系	比喻—拟人—夸张—排比—对偶—引用—设问—反问，并列短语—偏正短语—主谓短语— 动宾短语—后补短语—的字短语—介宾短语，举例子—打比方—作比较—列数字—分类别—下定义，举例论证—道理论证—比喻论证—对比论证，提出问题—分析问题—解决问题，小说—诗歌—戏剧—散文，论点—论据—论证，顺叙—倒叙—插叙—补叙，记叙—描写—说明—抒情—议论等
上下位关系	体裁(小说、诗歌、戏剧、散文)，表达方式(记叙、描写、说明、议论)，比喻(明喻、暗喻、隐喻)，论证手法(举例论证、道理论证、比喻论证、对比论证)，人物描写(肖像描写、语言描写、行动描写、心理描写、细节描写、神态描写)，复句类型(并列复句、转折复句、条件复句、递进复句、选择复句、因果复句、假设复句)，宋词(豪放派、婉约派)，句子(修辞手法、语法)，文章(体裁、中心思想、主题句，写作手法)，写作手法(象征、对比、衬托、烘托、照应、直接描写、间接描写、前后呼应、先抑后扬、先抑后扬)等
相反关系	写实—写意、拟人—拟物、正衬—反衬、对比—衬托、直接抒情—间接抒情、实词—虚词、立论—驳论、褒义—贬义、直接描写—间接描写、书面语—口头语，浪漫主义—现实主义等

3.非层次(对象属性)的推荐

本体中概念与概念的关系除了层次关系外，非层次关系也是本体的重要内容，如句子和修辞手法的关系，句子可以用到某种修辞手法，这是修辞手法可以作为一种关系连接句子和具体的修辞手法，即概念本身又可以作为一种关系，如图7所示。语文学科中有不少概念本身可以作为关系的例子，如字与拼音、文章与中心思想、文学作品与作家等、下位词作为上位词的关系。利用关联规则提取概念上下位关系后，同时也可以向本体工程师推荐是否把一个下位词作为上位词的一个关系。

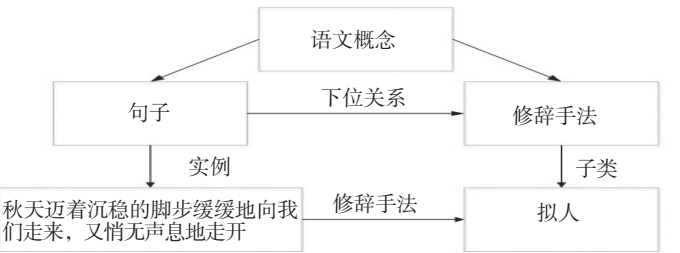

图7 修辞方法作为关系的例子

4.实例导入

目前互联网上语文学科已有大量的结构化知识，如在线字典、词典、古诗词、文章、作家等信息。本研究编写垂直爬虫，直接抓取结构化数据，利用JNEA结合构建好的骨架本体导致语义数据库中。算法流程如表4所示。

表4 实例导入算法流程

输入：领域专题网站
输出：实例三元组<Instance,rdf:type,Concept>，实例属性信息<Instance,Property,x>

关键步骤：
Step1确定要导入的概念Concept，获取概念对应的属性Property
Step2利用爬虫获取领域网站中包含实例的所有页面
Step3分析页面dom的模型，找到实例对应的节点
Step4获取到节点内容Instance，利用jena生成三元组<Instance,rdf:type,Concept>
Step5分析页面dom的模型，取得实例的属性Property对应数据x，利用jena生成三元组<Instance,Property,x>

5.人工修正

通过本体学习算法得到的本体很难做到完全准确，可以利用JENA的本体Validate方法对本体一致性进行效验(Pellet推理机在较大数据规模时会引发内存溢出无法使用)，最后再进行人工修正。学习元已开发出一个基于Web环境本体协同编辑系统，允许领域专家在Web形式下对本体进行修正，最终形成一个新的版本记录，从而保证本体构建的准确性，其算法流程如表5所示。

表5 人工修正流程

输入：本体效验请求
输出：修正和完善本体

关键步骤：
Step1运行Validate方法
Step2获得不一致的三元组Validity Report
Step3转到人工修正页面，请求领域专家修正
Step4领域专家修正并确认结果
Step5生成新的本体版本记录

此外，用户还通过检索相关语文知识，获得结果后对其进行改进，以成语“谨小慎微”为例，它属于词汇的子类，有同义词、反义词、近义词等关系，用户可以增加相关关系，也可以删除一些关系，经过管理员或领域专家审核后，也可以实现对本体的修正。

六、结论与后续研究

完善上述工作后，语文学科知识本体主要包含

数据如表6所示，知识实例数269012个，对应的三元组有2401244个。

表6 语文学科知识本体数据

概念	数量	三元组数	主要关系
汉字	21675	1194571	字义、同音字、反义字、近义字、拼音、声调、声母、韵母、笔画、偏旁等
词汇	153095	836848	词义、同义词、反义词、近义词、同音词、出处、情感维度、情感强度等
句子	17254	122545	出处、句类、句型、修辞、构成、主题词、语法、相关句子、前序句子、后续句子等
作家	894	13675	人物简介、朝代、国籍、出生年月、籍贯、代表作、作品、轶事典故、后世评价、主要成就等
古诗词	71893	206920	作者、标题、赏析、创作背景、翻译、诗意、相关诗词等
教材课文	1601	16007	文章类型(体裁)、作者、标题、赏析、创作背景、中心思想、主题句、表达方式(方法)、相关人物、相关情感、写作手法、后世影响、作品评价、关联文章等
人文性知识	2600	10678	相关人物、相关情感、关联价值观、相关名句、相关事件等
总计	269012	2401244	

中小学学科知识本体构建工作量巨大，本文提出了通过分析学科知识结构，梳理知识分类，结合领域专家和本体学习的方法实现学科知识本体构建，并以语文学科为例，验证框架思路的可行性，对中小学其他学科知识本体的构建有一定的启示作用。然而，本研究只实现语文学科中的部分知识本体，对知识实例的组织、语文学科教学等方面的本体构建还略显不足，需要进一步探索和完善。实际应用中，知识本体概念的实例和终端用户的联系更紧密。以语文中排比句这个概念为例，用户使用时可能希望得到各种实例，包括已经学过的课文中的排比句、课外名著中的排比句、即将要学习或已学习的排比句等，这些数据通过爬虫方式获取错误率较高，这就依赖于用户本体进化，实现实体的填充。语文呈现的表面知识如“冰山一角”，隐藏的则是整个社会民族的文化，这些人文性知识难以通过文本分析的手段直接获取。呼吁设计一种机制实现学科知识本体能够不断填充完善，使其具备进化功能，使得学科知识本体越来越智能。

参考文献：

[1] 余胜泉,杨现民,程罡.泛在学习环境中的学习资源设计与共享——“学习元”的理念与结构[J].开放教育研究,2009,(1):47-53.
[2] Welty C A, Jenkins J. Formal ontology for subject [J]. Data & Knowledge Engineering,1999,31(2):155-181.
[3] Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25-29.
[4] Angele J, Mönch E, Oppermann H, et al. Ontology-based query and answering in chemistry: Ontonova project halo[M].Heidelberg: Springer International Publishing, 2003.913-928.
[5] Dan McCreary. K-12 Education Metadata and the NIEMCase Study: Extending the National Information Exchange Model to Include K-12 Education[EB/OL].http://www.danmccreary.com/presentations/semweb2006/education_ontology.pdf,2015-1-10.
[6] 王晓琳,高丹丹,张际平.智能授导系统中的学习者本体构建[J].电化教育研究,2009,(4):56-77.
[7] 余莉,符红光.基于语义的几何学科知识平台[D].北京:中国科学院,2006.
[8] 詹川.基于教育心理学的课程知识本体模型研究[J].图书情报工作,2011,(14):111-115.
[9] 郝兴伟.基于知识本体的e-Learning系统研究[D].济南:山东大学,2007.
[10] 魏顺平.基于术语部件的领域本体自动构建方法研究——以教育技术学领域本体构建为例[J].电化教育研究,2013,(5):62-67.
[11] 顾芳.多学科领域本体设计方法的研究[D].北京:中国科学院研究生院 (计算技术研究所),2004.
[12] 杜小勇,马文峰.学科领域知识本体建设方法研究[J].图书情报工作,2005,(8):74-78.
[13] 赵呈领,黄志芳,万力勇等.基于初中物理课程的学科领域本体库构建研究[J].电化教育研究,2014,(8):64-94.
[14] 王冰洁.基于语义网的小学英语资源动态聚合系统设计与开发研究[D].北京:北京师范大学,2013.
[15] 郭军梅.基于初中语文学科本体的教学资源库的设计[D].长春:东北师范大学,2013.
[16] 李丽双.领域本体学习中术语及关系抽取方法的研究[D].大连:大连理工大学,2013.
[17] 徐建民,王金花,马伟瑜.利用本体关联度改进的TF-IDF特征词提取方法[J].情报科学,2011,(2):279-283.
[18] Chen N S, Wei C W, Chen H J. Mining e-Learning domain concept map from academic articles[J].Computers & Education,2008,50(3):1009-1021.
[19] Sanderson, Mark, Bruce Croft. Deriving concept hierarchies from text[DB/OL].http://delivery.acm.org/10.1145/320000/312679/p206-sanderson.pdf,2014-10-03.
[20] 张巍,于洋,游宏梁.面向词汇知识库自动构建的概念[J].现代图书情报技术,2009,(11):10-16.
[21] 邱福明.语文课程知识的存在论研究[D].济南:山东师范大学,2013.
[22] 韩雪屏.语文课程的知识内容[J].语文建设,2003,(3):4-6.
[23] 王荣生.语文科课程论建构[D].上海:华东师范大学,2003.
[24] 王荣生.“语文学科知识”概论——“语文学科知识精要”开篇语[J].语文学习,2011,(11):11-15.
[25] 李海林.“语文知识”:不能再回避的理论问题——兼评《中学语文“无效教学”批判》[J].人民教育,2006,(5):24-29.
[26] 邱福明.语文课程知识的存在论研究[D].济南:山东师范大学,2013.
[27] 周彩群.中学语文课程人文知识内容及其教学研究[D].长沙:湖南科技大学,2012.
[28] 中华人民共和国教育部.教育部关于印发义务教育语文等学科课程标准(2011年版)的通知[EB/OL]. http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/s8001/201404/167340.html,2014-08-01.
[29] 孙健.中文分词工具[EB/OL].http://www.nlpcn.org,2014-08-01.

(下转第124页)

作者简介：

于化龙：教授，硕士生导师，研究方向为信息化教学资源开发与应用(kjcyhl@mail.hebtu.edu.cn)。

崔杨：在读硕士，研究方向为信息化教学资源开发与应用(444634249@qq.com)。

The Reconstruction and Practice of Teaching Mode of *Technology of Modern Education* Experiment Course

Yu Hualong[1],Cui Yang[2]

(1.School of Education, Hebei Normal University, Shijiazhuang Hebei 050024;2.College of Information Technology, Hebei Normal University, Shijiazhuang Hebei 050024)

Abstract: The experiment course of "Technology of Modern Education" is an important curriculum of pre-service training for Teachers of the basic education. The promulgation of the New Criterion for Application Proficiency of Information Technology for Teachers of Primary and Secondary Schools in 2014 is a new requirement not only for primary and secondary school teachers in application proficiency of information technology, but also for students in normal colleges in training specifications of information technology. Under this background the paper, based on the existing problems in experiment courses of technology of modern education for the undergraduates in colleges, indicates reconstruction of the model of experiment teaching system for the Technology of Modern Education with the concept of applying micro video into the experiment. In the model, the course module is a foundation; the level of training is approaches; task-driven is a method; multivariate evaluation is an essential condition, and the four is closely related to form a complete teaching system model. Finally, according to this model from the perspective of teaching practice, a specific case is illustrated the process of applying model and its effect.

Keywords: Experiment for Technology of Modern Education; Teaching System;Level of Training; Task-driven; Proficiency of Educational Technology

收稿日期：2015年1月6日

责任编辑：宋灵青

(上接第89页)

作者简介：

丁国柱：在读博士，研究方向为知识本体技术、计算教育应用(laoding1982@qq.com)。

余胜泉：博士，教授，研究方向为教育技术基本理论、计算教育应用(toyusq@gmail.com)。

Research on Construction of Discipline Domain Ontology base on Skeleton Ontology and Ontology Learning

——Taking the Construction of Chinese Discipline Domain Ontology in Learning Cell as an Example

Ding Guozhu,Yu Shengquan

(School of Educational Technology, Beijing Normal University, Beijing 100875)

Abstract: Construction of Basic Education disciplines ontology is a huge amount of work, combined with domain experts and ontology learning is a viable method to build discipline ontology more efficient. In this paper we analyzed the Chinese discipline classification system, build a skeleton ontology; Meanwhile, taking advantage of TF-IDF ontology learning algorithm to extract key concepts and relationships in Chinese. Improve the Skeleton ontology of Chinese. Taking the knowledge of Chinese and the organization of the Chinese as a unity, eventually built more perfect Chinese discipline ontology in Learning Cell which supports experts to build ontologies manually and supports extracting key concepts and their relationships from the field in the domain document by machines. So that it can build a skeleton ontology faster and better.

Keywords: Ontology; Ontology Learning; Basic Education; Knowledge System

收稿日期：2015年1月11日

责任编辑：赵兴龙