文章编号:1006—9860(2013)11—0074—07

学习资源语义特征自动提取研究 *

杨现民[1]，余胜泉[2]

(1. 江苏师范大学 教育研究院，江苏 徐州 221116；2. 北京师范大学 现代教育技术研究所，北京 100875)

摘要：语义化学习资源的设计与建设已经成为e-Learning领域研究的热点。该文提出一种“语义基因”的概念来表征学习资源的语义特征，并对语义基因的自动提取方法进行了研究。语义基因反映了学习资源背后的内在知识结构，形式上表现为基于本体描述的带有权重的概念集合(包括核心概念以及概念间的关系)。语义基因在促进资源动态语义关联、资源检索、资源分类与聚类、资源进化等方面具有重要应用价值。实验表明，本研究提出的学习资源语义基因自动提取方法较之传统的词频法具有较高的召回率和准确率。

关键词：学习资源；语义特性；语义基因；领域本体

中图分类号：G434 **文献标识码：**A

一、引言

语义Web技术正在从实验室慢慢走向商用，各种语义化的产品和智能应用不断出现(如Powerset、Hakia、Twine)。语义Web的基本思想是用机器可处理的语义元数据描述Web资源，使得机器能对Web资源进行自动化处理，并智能地提供语义Web服务[1]。本体和推理作为语义Web体系架构的核心技术，引起了e-Learning领域研究者的极大关注，国内外众多研究机构和学者开始借助本体和推理技术来解决当前e-Learning领域存在的资源重复建设、检索效率低、个性化支持不足等问题。基于本体的教育资源具有权威性、规范性、可共享性等特点，借助本体技术可以促进数字教育资源的大范围共享，提升资源管理的效率和质量，实现资源的适应性推荐和递送。本体技术正在改变着数字化学习资源的组织方式，国内外出现了大量应用本体技术实现数字资源有效组织与管理的研究，主要集中在资源标注[2][3]、资源共享[4]、资源检索[5][6]、资源推荐[7][8]等方面。从维基到语义维基[9]，从学习对象到语义学习对象[10]，语义化学习资源的设计与建设已经成为e-Learning领域研究的热点。

当前大多数学习资源仍采用静态元数据的方式描述学习资源的语义信息，难以实现机器的自动理解和智能处理。学习资源是构建智慧学习环境的核心要素，而资源的语义化表征和组织是实现个性化资源推荐和适应性学习的重要前提。如何从语义层面表征学习资源的核心内容，如何实现学习资源语义特征的自动提取，是当前e-Learning领域学习资源进行语义化改造亟待解决的重要问题。学习资源的语义特征是指对能够表征学习资源核心内容的关键概念及其概念间的语义关系，采用本体技术进行语义化表示。本文在分析文本特征提取和语义特征提取方面已有研究的基础上，提出“语义基因”的概念，可用于描述资源的语义特征。本文对语义基因的概念进行了界定，提出一种学习资源语义基因的自动提取方法，并对提取的效果进行了初步检验。

二、相关研究

当前，有关资源语义特征提取的研究主要集中在语言学、图像视频处理、数据挖掘等领域。

语言学领域的语义特征分析与提取研究[11-13]的主要目的是从语义学的角度准确把握词汇、句法和语法所表达的语义信息。图像与视频语义特征提取研究[14-17]的主要目的是用于图像和视频的检索，即通过对高层语义特征信息的提取，提高检索的召回率和准确率。上述研究虽说是语义特征提取研究，但实际上在语义特征的表述方面并未涉及到领域本体库，即未采用规范的、语义化的方式表征资源的语义信息。目前，大多是将图像和视频中的信息通

* 本文系江苏省高校哲学社会科学基金项目“服务终身教育的泛在学习环境研究”(项目编号：2013SJB880033)和“移动学习”教育部—中国移动联合实验室开放课题“泛在学习资源的动态生成与协同进化机制研究”(项目编号：HX201307)的阶段性研究成果。

过文字形式(关键词等)表达出来，以供搜索引擎检索。

数据挖掘领域很多学者对文本资源的特征提取进行了研究。文本特征提取，即把从文本中提取出的特征词进行量化来表示文本信息，对文本进行科学的抽象，建立它的数学模型，用以描述和代替文本。当前有关中文文本特征提取的研究主要集中在特征项提取算法的设计上[18-20]，且大都基于统计学理论，没有结合领域知识，体现语义层面的需求和分析。

近年来，随着语义Web技术的发展，国内外已有少数研究者开始关注语义层面的文本特征提取[21-23]，即结合领域知识，采用概念词、同义词或本体来代替具体的关键词成为特征词，体现语义层面的需求。语义层面的文本特征提取方法较之纯统计学意义上的提取算法，理论上具有更高的准确率，提取的特征项更能全面反映文本内容的真实含义。不足在于，此类方法依赖领域本体库或主题词的建立，推广起来较为困难。

总的来说，上述有关语义特征提取的研究，处理的对象较为单一(视频、图像、文本等)，特征的表示上以“关键词”为主，缺乏规范的、语义化的描述。虽然数据挖掘领域的特征提取已经开始关注语义层面的需求，但仅用规范的概念替换以前的“关键词”，忽视了概念间语义关系的表征和提取。通过文献调研发现，当前数字化学习领域，学习资源的语义特征提取研究尚未引起研究者的重点关注。数字化学习资源语义特征信息的提取，对于提高资源检索效果、实现资源适应性推送、构建资源关系网络来说都具有重要价值。如何对具有复杂结构的数字化学习资源进行语义特征的表征和自动提取，是e-Learning领域急需解决的关键问题。

三、语义基因概念的提出

知识管理领域，刘惠植提出了知识基因的概念，认为知识基因是知识进化的最小功能单元，具有稳定性、遗传与变异性等特点，能够控制某一知识领域(学科、专业、研究方向)的发展走向[24]。吴力群指出，知识的基因是知识的内核，它由核心概念及核心概念之间的关系组成[25]。借鉴知识管理领域“知识基因”的概念，本研究将那些能够传达资源所要表达核心内容的概念及概念间的关系形象化地称之为“语义基因”，用于表征学习资源的语义特征。这里的语义基因是一种基于本体的资源语义特征表征方式，是本体的应用而非一个领域本体。

语义基因的应用价值主要体现在四个方面：(1)作为资源关联的计算数据，通过计算资源语义基因之间的语义关系(如相似、相反等)，动态建立资源之间的语义关联[26]，实现知识网络的自动生成与进化；(2)作为资源进化[27]的控制因子，通过计算资源语义基因和新增加内容的语义相似度，实现开放内容编辑的智能控制[28]，保证资源朝着预期的方向持续进化；(3)作为资源检索的重要信息源，通过对资源核心概念及其关系的检索，提高资源检索的召回率和准确率；(4)作为资源分类与聚类的重要依据，通过引入资源语义层面的特征信息，提高资源自动分类和聚类的速度和准确率。

1.概念界定

在界定语义基因的概念之前，首先对“语义”和“基因”的含义进行简要说明。语义学上的语义是指语言的意义，是语言形式所表达的内容。在计算机科学领域，语义是数据所表征的含义，是数据在某个领域上的解释和逻辑表示。基因的概念产生于遗传学，是控制性状的基本遗传单位。随后，基因的概念逐步渗透到文化学、管理学、计算机科学等多个领域。广义的基因概念是指能够决定和控制事物发展方向和表现特征的信息单元。

通过对“语义”和“基因”概念的分析，同时借鉴知识基因的定义，本研究中的语义基因的概念界定为：学习资源背后的内在知识结构，能够反映资源所要表达的核心内容。区别于文本相似度比较中的文档特征项，语义基因不是简单的关键词集合，而是资源背后所隐藏的语义概念网络。形象地说，语义基因就好比一棵大树的“树根”，控制着大树的性状和生长方向。

2.形式化定义

语义基因在形式上表现为基于本体描述的带有权重的概念集合(包括核心概念以及概念间的关系)。语义基因可以被形式化地表示为有序三元组，即SG = <CS, WS, RS>，如下页图1所示。其中CS是核心概念集合，CS = {C_1, C_2, C_3, … ,C_n}；WS是概念项的权重集合，WS = {W_1, W_2, W_3, …, W_n}，其中Wi为Ci的权重，$\sum_{i=1}^{n} W_i = 1$；RS为核心概念间的关系集，RS = {R_1, R_2, R_3, …, R_n}，每个关系采用领域本体中的RDF三元组<Subject, Predicate, Object>表示，R_1 =< Concept1, Relationship, Concept2 >，这里的Concept1和Concept2不一定包含在CS中，可以是领域本体库的其他概念，Relationship是从领域本体库中提取的概念关系。

为了更加清晰地解释什么是资源的语义基因，下面将以教育技术领域一段关于教学设计论述的文本(简称“教学设计论述”)作为例子，从中提取其

语义基因，如图2所示。需要说明的是语义基因的提取有特定的设计思路和实现算法(详见下文)，这里仅从形式上描述样例文本的语义基因：CS = {建构主义，教学设计，学习环境，自主学习策略，自主建构}，WS = { 0.35，0.25，0.1，0.15，0.15 }, RS = { <自主学习策略，下位概念，学习策略>，<建构主义，发展，认知主义>，<自主建构，互补，协同建构> }。

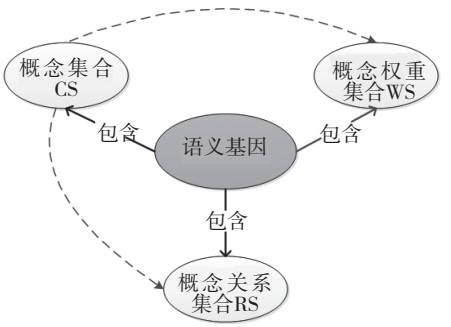

图 1 语义基因的结构要素

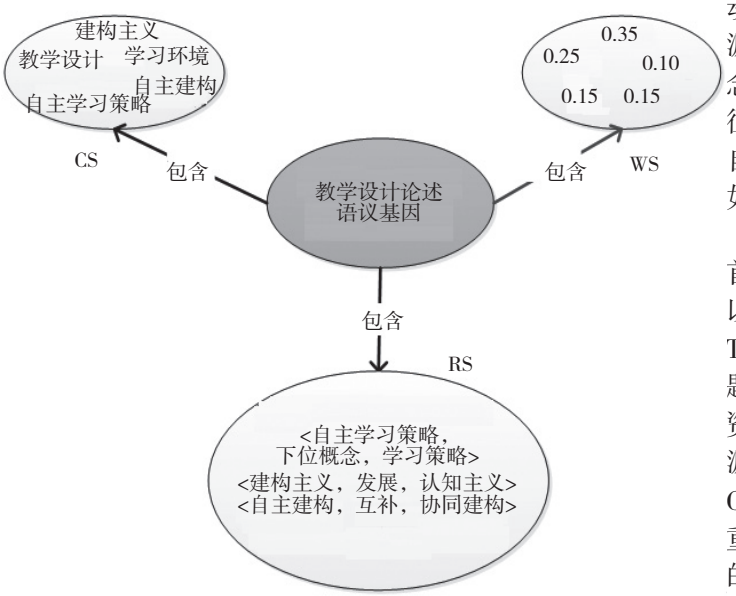

图 2 教学设计论述的语义基因

四、语义基因的自动提取

语义基因的自动化提取(Semantic Gene Extraction, SGE)类似Web数据挖掘中的文本特征提取(Text Feature Extraction, TFE)，都要从文本中提取最具代表性的文本特征，但又不同于TFE。TFE经常采用基于统计的方法提取文本中的关键词集，并通过构造评估函数计算特征词的权重，常用于文本的自动分类和聚类。SGE更加侧重于提取学习资源所传达知识的核心概念及概念间的关系，是语义层面的资源特征描述，而非统计学意义上的简单关键词集。SGE除了可以提高文本自动分类和聚类的准确度，还是实现学习资源动态语义关联的基础，通过语义特征词集合领域本体，可以计算出更加丰富的资源间的关系。此外，SGE还可以作为学习资源进化发展的“内在控制因子”，控制资源进化的方向。举个例子，一篇关于“建构主义教学设计”的文章，如果有用户试图将关于“一元二次方程解法”的内容加进去，该文章的“语义基因”便可以拒绝此次内容修改，从而在一定程度上控制资源的质量。

1.总体技术框架

提取学习资源语义基因的前提条件是领域本体库的建立，语义基因本质上是基于本体的资源内容特征项，即用标准化的本体数据来表征资源的核心内容。关于语义基因的设置主要有两种方式：一种是手动设置，即让资源的创建者手动添加语义基因，从领域本体库中选择能够准确表征资源内容的本体类，并赋予不同的权重；二是自动提取，即通过语义基因提取代理自动从资源的文本内容中提炼出核心的语义特征项(概念)及关系，并通过一定的规则为每个语义特征项赋予不同的权重。本研究重点研究的是自动化的语义基因提取方法，总体技术框架如图 3所示。

为了从学习资源的内容中提取语义基因，首先需要将资源实体进行结构化表征。这里可以将学习资源实体用四元组表示Res = <Title, Tag, Content, SemanticData>，Title表示资源的标题，Tag表示资源上附加的标签，Content表示资源的具体内容，SemanticData表示附加在资源上的基于本体的语义描述信息。Title、Tag、Content和SemanticData为语义基因提取的四种重要来源，在表征资源核心内容方面具有不同的重要程度。

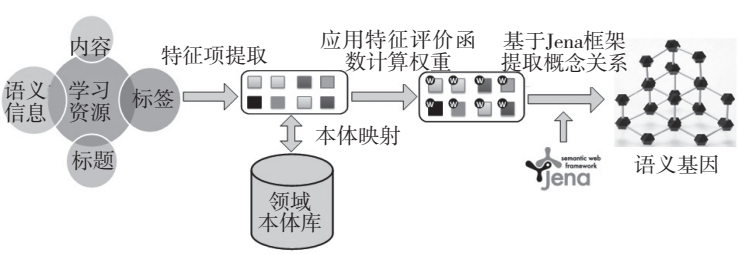

图 3 语义基因提取的总体技术框架

一般而言，资源的语义描述信息最为重要，SemanticData采用规范化的本体对资源内容进行描述，是获取语义基因非常重要的数据来源；其次，资源的标题也很重要，通过Title可以大体判断资源

的核心内容，用户在检索、选择资源时也常常依赖标题；再次，资源的标签是创建者为了从整体上描述资源而附加的特征词，常常也会成为用户判断资源内容和选择浏览资源的重要依据；最后，资源的内容是对资源的详细描述，由于数据丰富，承载了资源所要表达的核心内容，因此，也常常作为文本特征提取的重要来源。

本研究假设在语义基因提取方面，SemanticData所占权重大于Title所占权重，Title所占的权重大于Tag所占的权重，Tag所占的权重大Content所占的权重。权重集合可以表示为WT={WT_1, WT_2, WT_3, WT_4}，其中WT_1表示SemanticData所占权重，WT_2表示Title所占权重，WT_3表示Tag所占权重，WT_4表示Content所占权重。WT的初始值可以设置为WT = {0.4, 0.3, 0.2, 0.1}。

明确了语义基因提取四种重要数据来源及各自的权重后，接下来，借鉴Web数据挖掘领域较为成熟的文本特征项提取技术，同时结合领域本体库从资源中提取出一系列的特征词(核心概念)，并将这些特征词映射到本体，存放到CS集合中。然后，通过预先设定好的特征评价函数为每个特征项赋予不同的权重值，将这些权重值放到WS集合中。最后，通过Jena框架将这些特征词在领域本体库中存在的语义关系以三元组的形式提取出来放到RS集合中。

上述总体设计思路中包含四个关键性步骤，分别是基于领域本体的特征项提取，根据特征评价函数计算特征项的权重，特征词到本体概念的映射，基于JENA框架提取特征项(概念)在本体库中存在的语义关系。

2.基于领域本体的特征项提取

一般的文本特征项提取算法的主要步骤包括分词、停用词过滤、记录候选词在文献中的位置、根据TF-IDF计算词语权重、根据权重排序提取Top N的关键词等。语义层面的文本特征提取已成为文本特征提取的发展趋势。国内外已有少数研究者开始关注语义层面的文本特征提取[29-31]。语义基因的提取重在从语义层面提取能够表征资源内容的核心概念以及概念间的关系，而非简单的统计学意义上的关键词集合。

基于领域本体的特征项提取流程如图4所示。首先，将HTML文档中的html标签过滤掉，得到纯文本形式的资源内容。然后，调用中科院的中文分词工具ICTCLAS对文本内容和资源标题进行分词处理(标签已经是词集，不需要进行分词)，得到初始分词结果R_1。接着，将分词结果中的虚词过滤掉，只保留名词、动词和形容词，得到R_2。由于ICTCLAS分词的结果都是通用词典中的词汇，没有包含领域词汇，领域特征词在分词过程中会被切分成多个通用词汇，因此，需要结合领域本体将位置临近的通用词汇循环组合以识别领域特征词。比如ICTCLAS将“教学设计”切分成“教学”和“设计”两个词汇，通过结合教育技术领域本体可以将“教学设计”这一新的领域主题词识别出来，得到R_3。词语组合遵循“最长词语单元(Longest Term Unit，LTU)”原则，将几个相邻的能够准确表达某领域概念的词语组合成一个词语单元，因为一个LTU较之单个独立的词汇常常更能表达文本内容的核心思想。接下来，应用自然语言处理领域通用的中文停用词表过滤R_3中的停用词(所谓停用词就是在各种文档中经常出现的、不能反映文档内容特征的常用词，如：助词、语气词等)，得到R_4，并将资源附加的语义信息中提取的类(已是有意义的领域概念，不需要进行分词、虚词和停用词过滤)作为特征项合并到R_4中。由于R_4中可能存在多个同义词，为了保证特征项提取的一致性和权重计算的准确性，需要进行同义词替换，即将R_4中所有的同义词项替换为相同的词汇。这里使用哈工大扩展版的同义词词林[32](共77343条词语)结合领域本体进行同义词替换，替换的原则是尽量替换成本体中包含的概念，得到R_5。最后依据特征评价函数计算每个特征项的权重，得到带权重的特征项集合。

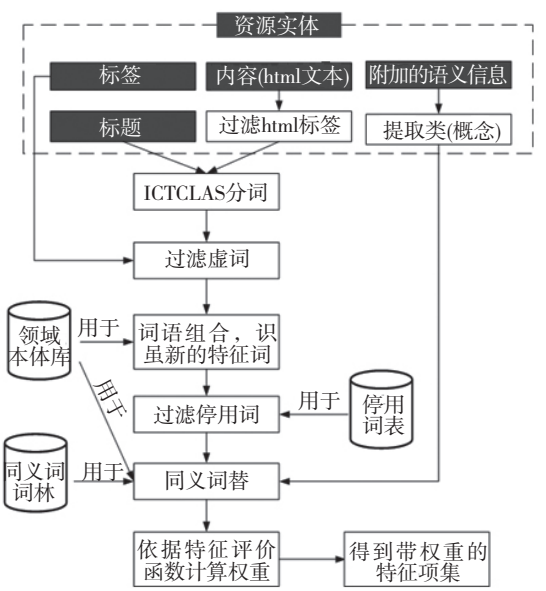

图4 基于领域本体的特征项提取流程

3.特征评价函数设计

TF-IDF(Term Frequency - Inverse Document Frequency)是计算特征项权重的经典理论，常用于

文本自动分类。TF-IDF是一种统计方法，用以评估一个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。TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。常用的TF-IDF公式为*TF-IDF=log(tf/df)*。

由于本研究不是针对文本自动分类进行的特征项提取，因此在特征评价函数的设计上可以不考虑DF(Document Frequency)，可以只以CF(Concept Frequency)作为权重计算的重要依据。结合上文对语义基因不同来源(标题、标签、内容和附加语义信息)的权重设计，语义特征项的特征评价函数可以表示为：

$$FE(t)=log(CF(c,SemanticDta)\times WT_1+CF(c,Title)\times WT_2+CF(c,Tag)\times WT_3+CF(c,Content)\times WT_4)$$

CF(c，x)表示概念c在x中出现的频度，$x\in\{SemanticData,Title,Tag,Content\}$。

4.特征词到本体概念映射

为了赋予上文得到的特征项规范化的语义信息，需要在特征词和领域本体的概念间进行映射(见图5)。假设特征词集为 $TS=\{t_i \mid i=1,2,3,...,n\}$，$CS=\{c_i \mid t_j \xrightarrow{map} c_i \wedge t_j \in TS, j\in[1,n]\}$ 映射成一个概念词集。

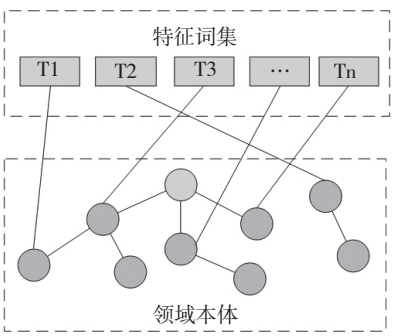

图5 特征词到本体概念的映射

本研究设计了基于JENA框架的特征词到概念的映射算法(Term Mapping to Concept，TM2C)。算法的输入项为资源的特征项集合TS，输出项为资源的概念词集合CS。算法的大体流程为：获取领域本体的JENA本体模型Model，用于操作本体；依次读取TS中的特征词，检查Model中是否包含该特征词，若包含则直接加入到CS中，如不包含则将该特征词生成本体概念加入到Model中，同时加入到CS中；返回本体概念集合CS。

5.概念关系提取

得到带权重的概念集合(可以分解为概念集合CS和权重集合WS)后，使用JENA框架编写概念关系提取算法，依次将此概念集中的概念项在领域本体库中存在的概念关系提取出来，通过三元组形式存放到RS集合中。

本研究设计了基于JENA框架的概念关系提取算法(Concept Relationship Extraction, CRE)。算法的输入项为资源的概念词集合CS，输出项为概念关系集合RS。算法的大体流程为：获取领域本体的JENA本体模型Model，用于操作本体；依次读取CS中的概念，从Model中遍历每个Statement，若Statement包含该概念并且Statement的Subject和Object都是概念，则将Statement加入到RS中；返回概念关系集合RS。

6.语义基因提取算法

结合上述语义基因提取的总体设计以及各关键步骤实现方法的分析，笔者提出如表1所示语义基因提取算法。

表1 语义基因提取算法

输入：资源的Title、Tag、Content和SemanticData
输出：资源的语义基因SG = <CS, WS, RS>
关键步骤：
Step1 调用ICTCLAS将Title进行分词处理和噪音过滤
Step2 调用ICTCLAS将Tag进行切割和噪音过滤
Step3 调用ICTCLAS对Content进行html标签过滤、分词处理、噪音过滤(去除虚词)
Step4 获取语义描述信息中的本体类
Step5 对Step2到Step5中得到的特征词集合，结合领域本体进行词语组合，识别新的特征词
Step6 调用停用词表，将Step6得到的词语集合进行停用词过滤
Step7 结合哈工大的扩展版同义词词林和领域本体进行同义词替换，得到特征词集TS
Step8 应用特征评价函数计算各特征词的权重，得到特征词的权重集合WS
Step9 应用TM2C算法得到概念集合CS
Step10 应用CRE算法提取概念关系集合RS
Step11 算法结束，输出CS、WS和RS

五、语义基因提取效果验证

本研究选择学习元平台(Learning Cell System, 以下简称LCS)[33]为实验环境，验证上述语义基因提取方法的效果。LCS是为泛在学习环境[34]设计开发的一种新型开放知识社区，官方网址为http://lcell.bnu.edu.cn。LCS以学习元作为基本的资源单元，学习元[35]是一种语义化组织的学习资源，多个学习元可以聚合成知识群。语义基因提取效果的评价标准采用召回率(Recall)、查准率(Precision)和$F_{measure}$指标。查准率用来衡量语义基因提取(不考虑权重)的准确性(精度)，即提取出来的正确的语义基因与提取的全部语义基因的百分比。召回率用来衡量语义基因提取(不考虑权重)的全面性，即提取出来的正确的语义基因与实际存在的语义基因总数的百分比。

$F_{measure}$是衡量语义基因提取整体效果(不考虑权重)的指标。

本研究采用将自动提取结果与专家人工提取结果相比较的方法验证上述语义基因提取方法的有效性。LCS中具备完善的教育技术领域本体库，采用随机方式选择15个学科属性为教育技术的学习元作为实验对象；选择一名教育技术学专家对学习元进行语义基因提取；然后，与系统自动提取的语义基因进行比较，统计Precision、Recall和$F_{measure}$指标值；分析比较结果，得出实验结论。

领域专家的选择满足两个必备条件：一是了解LCS平台，属于LCS的注册用户；二是在特定领域至少具有8年以上的研究经验，保证对上面选择的学习元的内容具有较深的专业理解。将待提取语义基因的学习元列表整理成Excel文件，通过e-Mail分别发给教育技术学科专家，并附加语义基因的解释信息和提取流程，便于专家正确理解、准确提取资源的语义基因。学科专家通过e-Mail返回语义基因提取结果，表2显示了部分系统提取和专家提取的语义基因(不含概念关系)。

表 2 系统提取与专家提取结果比较(部分数据)

	系统提取	专家提取
韦纳归因理论	归因0.155，归因理论0.111，失败0.111，学习动机0.102，原因0.1，行为0.097，成功期望0.086，学习0.083，能力0.08，情绪体验0.075	归因理论0.1，归因0.2，学习动机0.1，情绪体验0.1，成功期望0.1，后继行为0.1，积极归因0.15，消极归因0.15
信息技术与课程整合和CAI的区别	教学结构0.15，CAI0.145，整合0.13，学生0.124，信息技术课程0.098，教师0.093，区别0.078，信息技术0.062，知识0.062，计算机辅助教学0.057	信息技术0.2，课程整合0.2，CAI0.15，教学结构0.15，课程0.1，学生0.1，教师0.1
网络远程教育的特征	网络教育0.208，学习0.2，特征0.16，网络0.128，学历教育0.104，远程教育0.056，教育0.048，内容0.032，网络远程教育0.032，掌握0.032	网络远程教育0.15，网络教育0.2，特征0.15，内容0.1，学历教育0.1，掌握0.1，远程教育0.1，个性交流0.1

将专家返回的语义基因提取结果和系统自动提取的结果逐个比较，分别比较资源语义基因提取的召回率、准确率和$F_{measure}$的平均值。为了对比该方法的提取效果，采用文本挖掘领域传统的词频(Term Frequency)法对上述实验对象重新进行特征提取。两次提取结果比较见表3，这里将本研究提出的方法简称为“基因法”。

表 3 基因提取结果比较

	Recall	Precision	$F_{measure}$
基因法	0.87	0.61	0.71
词频法	0.76	0.50	0.60

实验结果表明，本研究提出的学习资源语义基因提取方法较之传统的词频法具有更高的召回率和准确率。虽然本研究提出的基因提取方法可以将资源实体包含的大部分核心概念提取出来，但在提取的准确率上略微偏低。接下来，还需进一步优化语义基因提取算法，提高资源基因提取的准确率。

六、结论与展望

学习资源的语义化组织是实现教育语义网的基础和前提。语义基因作为一种学习资源语义特征的表示方法，在促进资源动态语义关联、资源检索、资源分类与聚类、资源进化等方面具有重要应用价值。

本研究提出了语义基因的概念及其自动提取方法，并初步验证了语义基因提取的效果。不足之处在于：(1)语义基因提取方法的运行效率较低，占用过多的系统资源，目前在LCS中只能采用定时处理的策略，避免影响系统正常运行；(2)学习资源不会一成不变，改动后的资源需要重新提取语义基因，难以实现语义基因的实时更新。

本研究的后续工作将聚焦在三个方面：(1)优化语义基因提取算法，提升算法的运行效率，提高语义基因提取的召回率和准确率；(2)研究语义基因的进化问题，在资源变动的同时实现语义基因的实时更新；(3)依托LCS，探索语义基因的在资源检索、资源聚类、资源进化、资源关联等方面的具体应用，在实践过程中不断丰富和完善语义基因理论与方法。

参考文献：

[1] 李艳燕.基于语义的学习资源管理及利用[D].北京:中国科学院计算技术研究所,2005.

[2] 陈叶旺,李文,彭鑫,赵文耘.基于本体的文档语义标注改进算法[J].东南大学学报(自然科学版),2009,39(6):1109-1113.

[3] Weal, Mark J., Michaelides, Danius T., Page, Kevin R., De Roure, David C., Monger, Eloise and Gobbi, Mary. Semantic annotation of ubiquitous learning environments[J]. IEEE Transactions on Learning Technologies, 2012,5 (2): 143-156.

[4] 刘革平,赵嫦花.基于形式化本体的数字化学习资源共享技术研究[J].西南师范大学学报(自然科学版),2009,34(6):204-207.

[5] 郭广军,王剑波,游新娥,刘安丰.基于本体和语义网的网络教育资源检索研究[J].华中师范大学学报(自然科学版),2011,45(4):551-556.
[6] 涂军,曹鹏.数字图书馆中基于本体的语义检索模型研究[J].情报杂志,2012,31(7):191-194.
[7] 姜强,赵蔚,杜欣,梁明.基于用户模型的个性化本体学习资源推荐研究[J].中国电化教育,2010,(5):106-111.
[8] Ting-Peng Liang, Yung-Fang Yang, Deng-Neng Chen, & Yi-Cheng Ku. A semantic-expansion approach to personalized knowledge recommendation Original Research Article[J]. Decision Support Systems, 2008, (3): 401-412.
[9] Maged N. Kamel Boulos. Semantic Wikis: A Comprehensible Introduction with Examples from the Health Sciences[J]. Journal of Emerging Technologies in Web Intelligence,2009, (1): 94-96.
[10] Jesus Soto Carrion, Elisa Garcia Gordo, & Salvador Sanchez-Alonso. Semantic learning object repositories[J]. International Journal of Continuing Engineering Education and Life Long Learning, 2007, (17): 432-446.
[11] 邵敬敏,周芍.语义特征的界定与提取方法[J].外语教学与研究,2005,37(1):21-28.
[12] 杨帆.句法语义特征提取的认知语言学视角[J].语文学刊,2010,23(12):48-50.
[13] 陈珺.意欲形容词的语义特征分析[J].华南农业大学学报(社会科学版),2011,10(4):150-154.
[14] 贺莉娜.视频语义特征提取的研究[D].北京:北京交通大学,2008.
[15] Hyun-seok Min,Jae Young Choi,Wesley De Neve, &Yong Man Ro. Bimodal fusion of low-level visual features and high-level semantic features for near-duplicate video clip detection[J]. Signal Processing: Image Communication, 2011, 26(10): 612 - 627.
[16] 韩昌刚, 郭玉堂. 基于CCA的图像语义特征提取的分析与研究[J]. 计算机应用研究, 2012, 29(5): 1938-1942.
[17] Yin-Hsi Kuo, Wen-Huang Cheng, Member, IEEE, Hsuan-Tien Lin, Member, IEEE, and Winston H. Hsu. Unsupervised Semantic Feature Discovery for Image Object Retrieval and Tag Refinement[J]. IEEE Transactions on Multimedia, 2012, 14(4): 1079-1090.
[18] 尚文倩,黄厚宽,刘玉玲,林永民,瞿有利,董红斌.文本分类中基于基尼指数的特征选择算法研究[J].计算机研究与发展,2006,43(10):1688-1694.
[19] 张翔,周明全,耿国华.基于粗糙集的中文文本特征选择方法研究[J].计算机应用与软件,2010,27(3):4-7.
[20] 李凯齐,刁兴春,曹建军,李峰.基于改进蚁群算法的高精度文本特征选择方法[J].解放军理工大学学报(自然科学版),2010,11(6):634-639.
[21] Khan, A., Baharudin, B., & Khan, K. Semantic Based Features Selection and Weighting Method for Text Classification[DB/OL]. http://www.utp.edu.my/estcon2010/images/docs/itsim-final-approved.pdf,2013-09-09.
[22] Vicient, C., S á nchez, D., & Moreno, A. An automatic approach for ontology-based feature extraction from heterogeneous textual resources[J]. Engineering Applications of Artificial Intelligence, 2013, 26(3):1092-1106.
[23] 陈振亚,陈光辉,徐建民.一种基于本体的文本特征选取方法[J].广西师范大学学报(自然科学版),2011,29(1):143-146.
[24] 刘植惠.知识基因理论的由来、基本内容及发展[J].情报理论与实践,1998,21(2):71-76.
[25] 吴力群.知识基因、知识进化与知识服务[J].现代情报,2005,25(6):177-179.
[26] 杨现民,余胜泉,张芳.学习资源动态语义关联的设计与实现[J].中国电化教育,2013,(1):70-75.
[27] 杨现民,余胜泉.泛在学习环境下的学习资源进化模型构建[J].中国电化教育,2011,(9):80-86.
[28] 杨现民,余胜泉.开放环境下学习资源内容进化的智能控制研究[J].电化教育研究,2013,(9):83-88.
[29] Sukanya Ray, & Nidhi Chandra. Domain Based Ontology and Automated Text Categorization Based on Improved Term Frequency - Inverse Document Frequency[J]. I.J. Modern Education and Computer Science, 2012, (4): 28-35.
[30] Deng, Z. T., Hu, G. Y., Pan, Z. S., & Zhang, Y. Y. Kernel Sparse Feature Selection Based on Semantics in Text Classification[J]. Information Technology Journal, 2012, (11): 319-323.
[31] Khaled Elleithy. Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering[M]. Netherlands: Springer Netherlands, 2008.471-476.
[32] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform[DB/OL]. http://ir.hit.edu.cn/~car/papers/coling10demo.pdf,2013-09-09.
[33] 杨现民,余胜泉.学习元平台的设计开发及其应用场景分析[J].电化教育研究,2013,(3):55-61.
[34] 杨现民,余胜泉.生态学视角下的泛在学习环境设计[J]. 教育研究,2013,(3):103-110.
[35] 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,2009,15(1):47-53.

作者简介：

杨现民：博士，硕士生导师，讲师，研究方向为移动与泛在学习、数字化学习资源设计、网络教学平台开发、信息技术教学应用（yangxianmin8888@163.com）。

余胜泉：博士生导师，教授，主要研究方向为移动与泛在学习、教育信息化、信息技术与课程整合。

收稿日期：2013年3月12日

责任编辑：马小强