

中国现代职业教育理论之概念的抽取研究

王 良,梁 卿

摘 要:构建中国现代职业教育理论体系和概念体系的前提是确定中国现代职业教育理论的概念。基于术语学理论和实践,中国现代职业教育理论的概念获取摒弃了传统的纯人工方式,而是通过计算机技术的运用从专业文献中获得候选术语集合,从候选术语中筛选得到最终术语,并进而确定作为中国现代职业教育理论概念体系构成要素的概念。

关键词:现代职业教育理论;概念;术语

基金项目:国家社会科学基金“十二五”规划2013年度教育学一般课题“中国现代职业教育理论体系:概念、范畴与逻辑”(项目编号:BJA130096)

作者简介:王良,男,天津科技大学教育发展研究中心助理研究员,主要研究方向为教育理论与政策;梁卿,男,天津职业技术师范大学副教授,教育学博士,主要研究方向为职业教育原理与哲学,职业教育评价。

中图分类号:G710

文献标识码:A

文章编号:1674-7747(2016)19-0001-04

概念体系就是“根据概念间相互关系建立的结构化的概念的集合。”^[1]根据这一界定,概念是构成概念体系的要素。因此,构建概念体系首先需要识别并确定组成体系的概念。中国现代职业教育理论概念体系的建构亦不例外。但问题也随之而来,即如何才能识别并确定中国现代职业教育理论的概念呢?传统上,一般都是由专业领域的专家学者组成团队,采取人工的方式从专业文献中识别并收集概念。这种方式的典型代表就是各种专业词典的编纂。在人工方式中,概念的识别和收集是由专家学者集体完成的,因而具有很强的可靠性。但这种方式也具有明显的缺陷,即费时费力,效率低下。这一缺陷的进一步后果是,新出现的概念难以得到及时的识别和收集。如果采用这种方法收集中国现代职业教育理论的概念,并进而构建概念体系,那么,不仅概念的收录工作非常复杂,而且即使概念体系建构起来,其更新也将颇为缓慢。因此,从中国现代职业教育理论概念体系本身的发展来讲,需要找到一个新的获取概念的方法。

一、理论基础

我国国家标准《术语工作:原则与方法》(GB/T 10112-1999)指出,概念是人们根据客体(包括

客观存在且可感知的事物和想象产生的事物)的“特性概括而得到的心理构想。”“在一个知识领域中,概念用定义描述,被赋约定义的指称(即术语)。”^[2]国家标准的这些论述清晰地建构了一个“客体—概念—术语”的理论框架。根据这一框架,人们先是通过感官的观察和大脑的抽象,揭示客体的特征,并以此为依据对客体进行概括和分类,形成特定的心理构想,即概念。这是一个概念化的过程。作为一种心理构想,只有当概念能够被清晰地“说”出来的时候,才意味着概念的真正形成。因此,在概念化过程之后还有一个指称化的过程,即使用特定语词来代表或指称已经形成的概念。在这里,特定的语词就是术语。显然,客体先于概念,概念先于术语。

根据上文论述,在理想的情况下,概念的识别应与概念的形成是同一过程。这意味着,在中国现代职业教育理论概念体系的建构过程中,概念获取的根本途径就是通过对中国现代职业教育的现实进行观察和抽象。但这种途径存在巨大困难。第一,这个过程将会非常漫长。第二,在一个专业领域,一个所谓的“概念”是否称得上“概念”还需要得到同行的认可。也就是说,即使某个专家面向中国现代职业教育的现实抽象出了所谓

“概念”,还需要通过某种方式由同行进行认定。这又会进一步增加概念获取的难度。因此,试图在“客体—概念”的框架中来获得中国现代职业教育理论的概念并不具有现实的可行性。因此,我们只能在“概念—术语”的框架中来寻找办法。

事实上,在我国,学者们围绕着现代职业教育的各种问题,开展了大量研究,已经形成了一定数量的概念。在构建中国现代职业教育理论概念体系的过程中,人们的主要任务是从已有的关于现代职业教育的研究文献中,识别这些已经形成的概念。那么,如何才能识别已有概念呢?在“概念—术语”框架下,可行的路径是以术语识别为基础,获取概念。在理论上讲,术语只是表示概念的一种符号,获取了符号并不代表得到了概念。因此,试图通过术语识别的方式获取概念似乎并不可靠。但如果换个角度,这不失为一种比较好的方式。其一,从概念与术语的关系上看,概念虽然先于术语,但任何概念都必然表现为一定术语。其二,在术语学上,理想的情况是,一个特定概念在某一特定语种之中,一般只有一个相对应的术语。虽然现实并不如此,但这是术语学的努力方向。因此,在中国现代职业教育理论的概念体系建构过程中,人们可以通过对中国现代职业教育研究文献的考察与分析,从中提取相关术语,进而识别相关概念。

二、中国现代职业教育理论术语的收集方法

传统上,专业领域术语主要是由该领域的学者们直接从专业文献中识别。在信息技术尚不发达的情况下,这种方式即使存在缺陷,也是不可替代的。但是,近些年来,伴随着信息技术的发展和日渐成熟,人们开始尝试使用信息技术实现术语的自动发现和识别。从目前的实践来看,术语的自动识别一般有三种方法。

(一)基于规则的术语抽取法

这是一种以句法(术语上下文)或词法(术语的内部构成成分)为基础的术语抽取法。这种方法不仅需要对文档进行分词和词性标注,而且需要借助人工方式将语言学知识和特定专业领域知识结合起来,总结归纳特定专业领域术语构成知识,并用规则表达出来,进而使用计算机程序完成术语的自动抽取。基于规则的术语抽取法的最大优点是具有较高的准确性。但是,这种方法也具

有以下局限:(1)覆盖范围窄。即不同专业领域的术语的结构不同,因此,使用基于规则方法抽取术语时,每个专业领域都需要建立一套规则。(2)规则的建立费时费力。基于规则的抽取法的前提之一就是要建立规则,但规则的建立需要对大量专业文献进行研究才能完成,比较耗费时间。(3)不同时期术语的构成规则可能会发生变化。

(二)基于统计的术语抽取法

这种方法主要是基于术语区别于普通词汇的统计特征来实现术语的自动抽取。其优点在于:(1)简单直接。这种方法不需要借助规则等额外的要素,而是直接基于术语自身的统计特征来实现术语的自动识别,既简单又直接。(2)跨领域。基于统计的方法不用考虑特定专业领域的句法和词法等语言学信息,因此,这一方法适用于各个专业领域。(3)对于高频术语具有很好的识别率。其不足主要表现为:①受语言材料多少的影响较大。一般来说,在语料规模较大的情况下,这一方法的效果较好。②对低频术语的识别率较低。在专业文献中,有些词语出现的频率较低,但却是合法的专业术语。由于这类术语的出现频率低,在使用基于统计的方法识别术语时,这类术语往往会被错误地排除。

(三)混合策略的方法

将基于规则的方法和基于统计的方法结合起来使用。在实际的使用中,有的先采用统计方法,后采用规则方法,有的则相反。^[1]

在上述三种方法中,理论上比较理想的方法是混合策略的方法。但有研究表明,这种方法并不是在所有情况下都是最优的选择。^[4]本研究拟将基于统计的自动抽取法和人工方法结合起来使用。具体而言,本研究首先采用基于统计的自动抽取法中的信息熵方法形成初步的术语集合。然后采用人工的方式,对初步的术语集合进行分析,形成最终的专业术语。之所以采用基于统计的方法而不是基于规则的方法有两点原因:一是基于规则的方法需要建立规则,而规则的建立比较复杂。职业教育领域术语规则的建立需要语言学家和职业教育学者的合作,需要耗费很长的时间。借助已有的相关词典(顾明远教授主编的《教育大辞典》第3卷中收录有职业教育方面的术语)来建立规则虽然也是可行的路径,但该词典出版于20

多年前,职业教育术语的构词规则有可能发生变化。二是因为统计方法的优缺点符合我国职业教育研究的特点。近些年来,我国职业教育研究水平有了长足的进步,但总的来说存在理论水平不高,理论色彩不够浓厚的问题。许多研究成果往往就事论事,满足于对职业教育现状的简单描述。^[5]可以说,相对于其他领域,包括普通教育和高等教育领域而言,我国职业教育研究水平相对落后。这意味着,在职业教育理论中,不仅概念的科学性还有待提高,而且科学的概念在数量上相对较少。一般来讲,科学的、得到公认的科学概念使用的频率相对比较多。基于统计的术语自动抽取法,高频术语的识别较强,低频术语的处理较弱,这个特点与目前职业教育的研究状况具有很强的适切性。同时,从促进职业教育理论发展的角度,在职业教育研究总体水平有待进一步提高的条件下,不能急于提取尚未成熟的概念,而更应该将已经得到认可的、被学者们广为使用的术语识别并确定下来。之所以采用信息熵方法,主要是因为它是目前被广泛认可并使用的信息统计工具。之所以在基于统计方法之后使用人工方法进行进一步分析,主要是因为,统计方法在术语抽取中存在一大特点,即它不仅能识别文献中的高频术语,还往往会将高频通用词汇和其他字符串抽取出来。这些词语如果不通过规则的方式予以排除,就需要借助人工的方式去除。

三、中国现代职业教育理论术语的收集和概念的识别

在本研究中,“中国现代职业教育”被界定为中华人民共和国建国后的职业教育。因此,在术语收集过程中,应该将新中国建国以来的职业教育研究文献都囊括进来。但这样一来,工作量将非常巨大。事实上,也完全没有必要这样做。20世纪80年代末90年代初,北京师范大学顾明远教授主持编纂出版了《教育大辞典》(12卷),其中第3卷收录了职业教育方面的术语。我们将该词典收录的413条术语全部纳入进来(不含人名、具体的学校名、法律规范名、会议等专有名词),作为中国现代职业教育理论术语的一部分。^[6]如此以来,我们需要收集处理的就是20世纪90年代以来的术语。具体过程如下。

(一)建立语料库

以“职业教育”或“职业技术教育”为主题,在中国知网期刊网CSSCI库搜索职业教育领域2014年(含)之前的研究文献(不包含研究综述、新闻报道、会议综述、领导讲话和政府文件类文章),下载这些文章的题名、关键词和摘要,以文本文件形式存储,得到由3 083个文本文件构成,共计60余万字的语料库。

(二)二次开发NLPIR分词系统

NLPIR分词系统是由汉语词法分析系统ICT-CLA发展而来。该系统曾获得钱伟长中文信息处理科学技术一等奖等多个奖项,在中文信息处理领域被广泛认可。目前,该系统在传统的分词和词性标注功能之外,又相继开发出了许多新的功能。基于信息熵方法的关键词自动获取就是其新的功能之一。基于信息熵方法是基于统计方法中的一种,其基本思想是,如果一个词或短语是术语或者关键词,那么它反复出现并且前后出现不同词语的概率应该非常高。作为NLPIR系统的一部分,该系统的关键词提取组件采用了灵活的开发结构,可以支持各种不同的计算机操作系统和调用语言。本研究在win32系统中,采用Java语言,对NLPIR系统进行了二次开发,以实现基于信息熵方法的术语自动获取。

(三)术语的自动提取、选择与校对

运行术语提取程序,处理3 083条题录信息,初步获得候选术语24 385条。在此基础上,采取人工的方式对这些术语进行选择 and 校对,最后获得职业教育领域的专业术语349条。术语的选择和校对遵循“四排除、一合并”原则。“四排除”是指:(1)高频非通用词汇予以排除。例如,“发展残疾人”就不是术语。它之所以被抽取出来,主要是因为结合比较稳定而且出现次数多。(2)通用词汇予以排除。如“原因”、“启示”、“探索”等通用词汇被排除在术语之外。(3)专有名词予以排除。例如,“黄炎培”、“杜威”、“普洛瑟”等人名、“澳大利亚”、“江苏”、“株洲市”等地名,“浙江金融职业学院”、“中华职业学校”等特定学校的校名等都被去除。(4)未体现职业教育特色的教育术语予以排除。例如,“教育”、“课程”、“教学”等虽然是教育术语,但并未体现职业教育特点,因此,也被排除在职业教育术语之外。“一合并”是指将在不同文

献中出现的同一术语加以合并。

(四)术语的确定

将《教育大辞典》(第3卷)中收录的413个职业教育术语与运用信息熵方法提取的349个术语进行比对,取其并集,形成中国现代职业教育理论的术语集合,共计733个术语。两个来源的术语集共有29个重复术语。应该说,重复的术语数量比较少。出现这种状况的可能原因是,CSSCI期刊即使接收职业教育方面的论文,一般也不刊发教学方面的文章。这就使得基于CSSCI期刊库获得的职业教育术语很少涉及职业教育微观教学领域。但《教育大辞典》(第3卷)收录的职业教育术语中,涉及职业教育微观教学领域和相关职业领域的占比非常高,前者如“案例教学法”、“示范教学法”等,后者如“工种”、“成品测量”等。

(五)概念的确认

虽然术语学的追求是一个术语对应一个概念,但在我们收集获得的733个职业教育术语却不是这样。在这733个术语中,有些术语虽然不同,但表达的实际上是同一个概念。例如,“职业技术教育”、“职业教育”、“技术和职业教育”和“技术职业教育”等四个术语所指的其实是一个概念。“职业基础教育”与“基础职业教育”、“职业教育体系”与“职业技术教育体系”也是如此。对于这些术语,我们将其进行归并,最终形成了中国现代职业教育理论的728个概念。

本研究也存在一些不足。其一,限于能力和时间,在语料库的建设上,本研究只是将中国知网CSSCI期刊库论文的题录信息收集起来,既没有收集更多论文的题录信息,也忽视了20世纪90年代以来出版的职业教育著作。其二,在术语自动抽

取方法上,本研究采取了信息熵方法。这种方法虽然考虑了一个词或短语外部的边界自由度,即词或短语左右两侧出现不同符号或符号串的可能性大小,但没有考量词或短语内部结合的紧密度,即一个词或短语内部各符号间的结合强度。^[4]通常情况下,作为概念的指称,术语应是一个相对固定的完整语言单位,不仅其左右两侧的符号和符号串应该是变化的,即边界自由度高,而且其内部符号之间也应该紧密结合在一起,即内部紧密度高。信息熵方法只考虑了边界自由度,而忽视了内部紧密度,这不能不说是一大缺憾。在后续的研究中,我们考虑将职业教育研究领域主流期刊论文的题录信息涵盖进来,以扩大语料库规模,并将尽可能采取措施改进术语自动抽取的方法,以对这些缺憾加以弥补。

参考文献:

- [1] 中华人民共和国国家质量监督检验检疫总局.术语工作:概念体系的建立(GB/T 19100-2003)[Z].北京:中国标准出版社,2003:1.
- [2] 国家质量技术监督局.术语工作:原则与方法(GB/T 10112-1999)[Z].北京:中国标准出版社,2000:1.
- [3] 季培培,鄢小燕,岑咏华.面向领域中文文本信息处理的术语识别与抽取研究综述[J].图书馆情报工作,2010(16):124-129.
- [4] 张二艳.术语自动抽取技术研究[D].哈尔滨:哈尔滨工业大学,2009:9,18.
- [5] 徐国庆.职业教育原理[M].上海:上海教育出版社,2007:12-14.
- [6] 教育大辞典编纂委员会.教育大辞典(第3卷)[Z].上海:上海教育出版社,1991:227-341.

[责任编辑 曹 稳]