课程本体自动构建技术研究

童名文[1] 牛 琳[1] 杨 琳[1] 邹军华[2] 上超望[1]
(华中师范大学教育信息技术学院 武汉 430079)[1] (湖北大学教育学院 武汉 430415)[2]

摘 要 课程本体是课程知识组织的一种重要技术,在智能学习系统中得到广泛应用。针对人工建立课程本体依赖专家经验和效率较低等问题,提出课程本体自动构建技术。该技术以丰富的 Web 课程资源为数据源,集成网络爬虫、中文分词和关联规则挖掘等技术,实现课程本体自动构建。实验结果表明,该技术建立的课程本体不仅具有较好的质量,而且执行效率较高。

关键词 课程本体,本体自动构建,中文分词,关联规则挖掘,网络爬虫
中图法分类号 TP37 **文献标识码** A

Research on Technique of Course Ontology Automatically Constructing

TONG Ming-wen[1] NIU Lin[1] YANG Lin[1] ZOU Jun-hua[2] SHANG Chao-wang[1]
(School of Education Information Technology, Central China Normal University, Wuhan 430079, China)[1]
(School of Education, Hubei University, Wuhan 430415, China)[2]

Abstract The course ontology used in the artificial intelligent learning system widely is one of the important technologies for knowledge organization in courses. To solve the problems of manual technique to construct the course ontology, a novel technique was proposed to construct the course ontology automatically. The technique employes some techniques such as Web spider, Chinese character division, and relation rule mining, to extract the concepts and relations of the course ontology from diverse web course resources. By experimental evaluation, it is proved that the technique can construct the course ontology automatically with high quality and efficiency.

Keywords Course ontology, Ontology automatically constructing, Chinese character division, Relation rule mining, Web spider

1 引言

自 20 世纪 70 年代起,知识系统就已成为人工智能领域研究的热点之一。知识库作为知识系统的核心组成部分,一直受到研究人员的广泛关注。知识库研究主要涉及两方面问题:1)知识的组织、表示与构建理论和技术;2)知识的推理机制。关于知识的组织与表示已经有诸多理论和技术,其中本体是知识的一种重要组织技术,它通过概念、属性、实例以及它们之间的关系来显式表达某一领域的知识。

本体构建技术主要研究如何建立课程本体的解决方案。目前,本体构建技术主要分为人工构建和自动构建两类。人工构建是依靠专家经验建立本体。人工建立本体的优势在于精细、准确度高。但构建过程需要领域专家和技术人员共同参与完成,会耗费大量的人力。自动构建是利用客观数据建立本体。本体自动构建技术中存在两个关键问题:1)如何获取与领域知识相关的大量客观数据;2)如何消解概念歧义[1]。由于这两个问题一直没能得到很好的解决,因此在本体构建技术研究中人工方式一直占据主导地位。然而,在课程本体自动构建技术研究中,上述两个问题已经能够解决。其一,随着网络学习的兴起,各类网络课程数量迅速增长。据统计,截止 2013 年,我国国内网络课程数已多达 10 万门(包括精品课程、精品资源共享课和视频公开课等),这些网络课程为课程自动构建提供了丰富的数据源。其二,在一门课程中概念实质就是课程术语,而课程的核心术语是统一的,出现歧义的情况很少,所以,课程本体自动构建技术可以回避语义消解问题。因此,课程本体自动构建技术的两个关键问题可以得到很好的解决,技术实现具有可行性,然而国内外关于课程本体的自动构建技术研究还较少见诸于文献。

本文以网络课程资源为数据源,基于网络爬虫、中文分词和关联规则挖掘等技术,提出了一种课程本体自动构建技术。该技术将课程本体自动构建过程分为资源爬取、领域概念提取、本体关系挖掘和本体描述 4 个步骤。首先制定资源的筛选策略和网络爬虫的爬取规则,自动获取课程相关数据;其次采用中科院分词模块(imdict-chinese-analyzer)和 Nutch 分词

本文受教育部人文社科基金资助项目:数字化学习资源无障碍适配决策模型研究(15YJA880062),中央高校基本科研业务费项目:内容适配系统中最优适配决策器模型及分布式寻优算法研究(CCNU14A02012)资助。
童名文(1975—),男,博士,教授,主要研究方向为多媒体内容适配技术、知识建模;牛 琳(1988—),女,硕士生,主要研究方向为知识建模;杨 琳(1974—),女,博士生,讲师,主要研究方向为信息技术教学应用,E-mail:50944299@qq.com(通信作者);邹军华(1972—),男,博士,副教授,主要研究方向为信息技术教育应用;上超望(1980—),男,博士,副教授,主要研究方向为数字版权保护。

工具提取课程词汇，建立关键词的索引表，计算关键词的出现频率，将频率大于阈值的词汇作为课程本体概念；然后，通过关联规则挖掘技术在数据源中做概念直接的关联分析，计算关联强度，将强度大于阈值的关联作为课程本体关系；最后在自动提取的课程概念和关系基础上，人工增加属性和实例，并用本体描述工具建立课程本体，完成课程本体的构建。

2 相关研究

本体自动构建技术采用计算机为主、人工辅助的方式实现本体构建。近 10 年，已有一些本体自动构建的技术，如 Microsoft 公司的商业产品 MindNet 以词典和百科全书为数据源，自动建立领域本体；Paola Velardi 等人开发的 OntoLearn的文本挖掘工具可以在文本数据中自动挖掘概念和关系，自动生成本体；Harith Alani 提出了重用已有本体，自动构建生成新的本体的技术[2]。按照数据来源的不同，国内外关于本体自动构建技术的研究大体可以分为 3 种类型。

2.1 基于语义词典的本体构建技术

有别于传统的词典，语义词典不仅有词汇，还包含词汇之间的关联，这很适合本体的自动构建。当前已有一些通用的语义词典，如 WordNet 和 HowNet，它们可以用于通用本体的构建。对于领域本体构建，需要领域语义字典，如《医学主题词表》[3]和《化学工业主题词表》等。虽然基于语义词典的本体自动构建技术实现相对简单，但语义词典尤其是领域语义词典较少。数据源缺乏限制了对该技术的研究。

2.2 基于文本的本体构建技术

基于文本的本体构建技术以文本为数据源，采用文本挖掘技术建立本体。常用的文本挖掘技术有模板识别、关联规则、概念聚类、概念学习等[1,4]。在国内，刘磊等提出了基于模板识别的 SSE_CMM 领域本体自动构建技术[5]。国外，Agrawal 等提出了通过关联规则挖掘建立本体的技术[6]；Hearst 等提出了基于文本挖掘发现知识的技术[7]；Joachims 提出将 TFIDF 文本聚类算法用于本体构建的技术[8]；Kietz 等提出了基于概念学习的本体构建技术[9]。

2.3 基于知识库的本体构建技术

知识库是结构化的知识集合，能够实现知识的机器可读和高效管理。基于知识库可以很便利地提取本体的概念和关系[10,11]。陆文豪等提出了一种基于关系数据库的本体构建技术[12]，但这种技术依赖于关系型知识库，难以延伸到其它类型的知识库。随着技术的发展，知识库的形式不断更新，越来越多的研究人员开始关注基于百度百科、维基百科等开放知识库的本体构建技术[13]。但由于这类知识库的异构性和非结构化，此类知识库的本体构建技术研究还处于起步阶段。

3 技术方案设计

课程本体自动构建技术以 Web 网络课程为数据源，将课程本体构建过程分为 4 个步骤，如图 1 所示。第一步，利用网络爬虫(spider)自动抓取 Web 课程网页资源作为课程本体构建的数据源；第二步，对抓取的 Web 文本资源做中文分词，并建立关键词倒排索引，通过分析倒排索引的高频词汇，自动获取课程本体中的概念；第三步，在 Web 课程网页资源中做概念关联挖掘，提取课程本体中的关系；最后，将提取的概念和关系加以整理，用本体建模工具描述课程本体，完成课程本体构建。

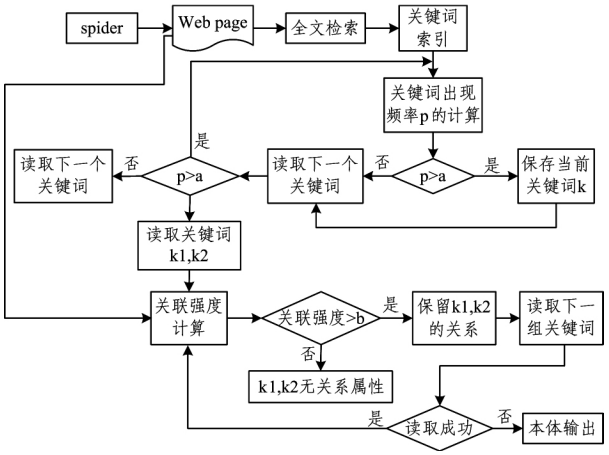


图 1 课程本体自动构建流程

在课程本体自动构建技术中的关键步骤是中文分词、概念提取和关系发现。下面简要介绍这 3 个步骤的技术实现。

3.1 中文分词

目前已有一些较好的英文分词工具，如 Nutch。但因为中英词汇的不同，分词工具在中文和英文的分词上性能差异很大。Nutch 在英文分词中具有良好的性能，但在中文分词方面表现不尽如人意[14]。针对中文分词需求，中国科学院计算技术研究所提出了中文分析器——imdict-chinese-analyzer。该分析器采用层叠隐马尔科夫模型将分词算法、切分排歧和未登录词识别有机地融合在一起，能够有效地实现中文词汇分离，实验测试表明该技术在中文分词上的性能优于 Nutch[15]。因此，中文分词模块采用 imdict-chinese-analyzer 中文分析器实现。

3.2 概念提取

概念提取的基本原理是课程本体中的核心概念在课程资源中出现的概率较大。因此，将在课程资源中出现频率高的词汇作为课程本体的概念。概念提取中包括两个主要问题：词汇出现概率计算和概率较大表示。第一个问题采用频率近似概率的方法解决。通过网络爬虫自动抓取一定数量的 Web 课程网页资源(不少于 300 个页面)，采用全文检索技术为每个网页建立关键词索引，将所有关键词索引融合为一张整体的关键词索引表，最后统计每个词汇出现的频率，将频率值作为每个词汇出现的概率。频率计算如式(1)所示：

$$p(w_i)=\frac{sw_i}{sw}*\lambda \tag{1}$$

其中，sw_i 是第 i 个关键词出现的次数；sw 是所有关键词出现的总次数；λ 为标量参数，用于调节频率分布范围。

对于第二个问题，通过阈值来描述概率较大这一语义，当词汇出现概率大于阈值时，词汇作为课程本体的概念，否则词汇不进入课程本体。

3.3 关系发现

课程本体关系发现的基本原理与概念提取相似，仍是本体中的关系在课程资源中出现强度较大。具体过程是，在课程资源中，利用关联规则挖掘技术抽取概念之间的关联，然后计算每个关联的频率，将其作为关联的出现强度。关联频率的计算如式(2)所示：

$$rm(r_{ij})=\frac{sr_{ij}}{sr_i+sr_j}*\theta \quad (2)$$

其中，sr_{ij} 是出现关键词 i 和关键词 j 关联的次数；sr_i 是包含关键词 i 的关联次数；sr_j 是包含关键词 j 的关联次数；θ 为标量参数，用于调节关联强度的分布范围。

通过阈值表达强度较大，当强度大于阈值时保留关联作为本体中的关系，否则放弃该关联。

4 技术实现与评价

为检验所设计技术方案的有效性和技术性能，采用 Java 技术实现了课程本体自动构建技术方案，并以《C 语言程序设计》课程为例，使用该技术建立《C 语言程序设计》课程本体，通过实验对技术做功能和性能评估。

4.1 技术实现

课程本体自动构建技术方案的实现工具及功能描述如表 1 所列。下面以《C 语言程序设计》课程为例，从资源爬取及索引建立、领域概念提取、关系发现以及本体建立 4 个方面分别简要描述技术的执行过程及结果，以证明技术的有效性。

表 1 开发工具及功能描述

开发工具	功能描述
MyEclipse9	Java 集成开发环境，用于辅助完成代码的编辑以及调试任务
Nutch1.0	基于 Lucene 的 Java 应用系统，可实现搜索引擎应用，本实验中用于完成网页爬取和下载的功能
imdict-chinese-analyzer[1]	中科院研发的中文分词插件，优化 nutch1.0 的中文分词，为后续索引的建立、概念和实例的抽取做准备工作
JavaCC6.0[2]	Java 开发的语法分析生成器，可编译文件生成 Java 程序，实验中用于编译修改后的分词算法
Luke3.5[3]	Andrzej Bialecki 开发的一个索引工具箱，用于查看索引文件的内部内容以及完成一些特定的查询操作
Protégé4.3	基于 Java 语言的本体开发工具，用于本体自动构建完成后的本体检验以及人工调整

(1)资源爬取与索引建立

选取《C 语言程序设计》课程中的核心词汇“C 程序设计”、“数据类型”、“分支结构”、“循环结构”、“数组”、“函数”、“预处理”、“宏定义”、“指针”、“结构体”、“共用体”、“位运算”和“文件”作为关键词列表文件 words.txt，得到 url 种子集文件。设置爬虫的爬取规则：最大爬行深度为 3 层，下载页面上限为 1000，启动爬虫并基于中文分词完成索引的建立，运行结果如图 2 所示。

Top ranking terms. (Right-click for more options)

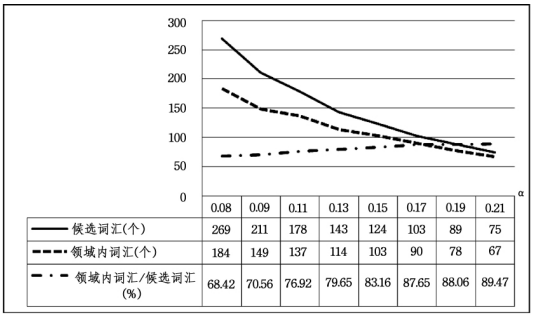
No	Rank	Field	Text
1	834	url	http
2	510	url	www
3	510	host	www
4	391	content	问题
5	376	content	程序
6	363	content	设计
7	362	content	技术
8	347	content	语言
9	335	content	开发
10	315	content	reserv
11	307	content	信息
12	305	url	html
13	303	content	管理

图 2 Luke 建立索引的结果

实验中共爬取 834 个网页，包含 32718 个候选词组，其中正文部分(content)的词组有 30390 个，占总词组的92.88%，因此后续的概念提取和关系建立均基于正文部分(content)展开分析。

(2) 概念提取模块

根据式(1)对建立的索引进行统计分析。由图 2 可知，数量较多的索引项决定了所有关键词出现的总次数较大，求得的概率值较小，因此为便于观察实验数据，对概率分布范围进行调整。设置标量参数 $\lambda=100$，即概率值放大 100 倍，此时概率区间分布在 0 到 2 之间。经过多次实验测试，概率阈值 α 与候选词汇、领域内词汇以及领域内词汇占候选词汇的比值关系如图 3 所示。由图 3 可以看出，随着概率阈值 α 的逐渐增大，候选词汇和领域内词汇的个数逐渐减少，但领域内词汇占候选词汇的比值逐渐增大，即概念提取的准确性逐渐加强。



α	0.08	0.09	0.11	0.13	0.15	0.17	0.19	0.21
候选词汇(个)	269	211	178	143	124	103	89	75
领域内词汇(个)	184	149	137	114	103	90	78	67
领域内词汇/候选词汇(%)	68.42	70.56	76.92	79.65	83.16	87.65	88.06	89.47

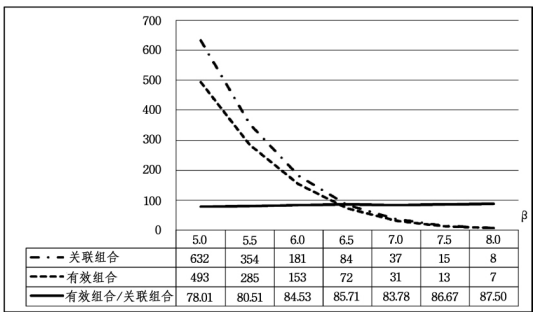
图 3 α 值对概念提取的影响

当概率阈值 α 取 0.15 时，共有 124 个候选词汇，其中领域内词汇 103 个，占候选词汇的 83.16%，以此为例，运行结果如图 4 所示。

关键词	文档数	词频	概率
content:0000	1	105	0.255
content:16	16	64	0.155
content:ch	9	88	0.213
content:char	21	197	0.479
content:file	7	83	0.201
content:for	21	79	0.192
content:fp	4	181	0.440
content:if	23	92	0.223
content:includ	17	65	0.158
content:int	24	397	0.965
content:long	9	62	0.150
content:main	20	78	0.189
content:name	9	65	0.158
content:printf	25	185	0.449
content:ptr	5	134	0.325
content:sizeof	15	108	0.262
content:struct	11	85	0.206
content:unsign	6	67	0.162
content:while	19	102	0.248

图 4 《C 语言程序设计》课程本体概念结果

以概念提取模块的运行结果为基础，在筛选出的词汇之间建立关系。在概率阈值 α 取 0.15 的运行结果基础上，设置标量参数为 $\theta=10$，使关联强度分布在 1 到 10 的区间内。依据式(2)进行关联强度计算，经实验测试，词汇关系随关联概率阈值 β 的变化如图 5 所示。



β	5.0	5.5	6.0	6.5	7.0	7.5	8.0
关联组合	632	354	181	84	37	15	8
有效组合	493	285	153	72	31	13	7
有效组合/关联组合	78.01	80.51	84.53	85.71	83.78	86.67	87.50

图 5 β 值对关系发现的影响

图 5 中的关联组合是指由式(2)计算出的达到阈值要求的概念对总数，有效组合是在关联组合的基础上，经人工处理筛选出的有意义的概念对总数，有效组合/关联组合旨在通过

计算关联组合中的有效比例反映关系建立的准确性。可以看出，在 5.0 到 8.0 的变化区间内，随着关联强度阈值 β 的增大，有效组合和关联组合的组合数逐渐下降，但有效组合与关联组合的比值相对稳定，基本保持在 80%～85% 左后，说明由此得出的词汇关系具有较高的准确性。由实验可知，当关联强度阈值 β 为 6.5 时，共得到 84 对关联组合，其中有效组合 72 对，占关联组合的 85.71%，以此为例，运行结果如图 6 所示。

知识关系结构如下所示：

关键词1	关键词2	关联强度
content:char	content:int	8.749
content:char	content:内存	6.8
content:char	content:字节	6.666
content:if	content:语句	6.562
content:if	content:输入	7.142
content:includ	content:main	7.619
content:int	content:printf	6.896
content:int	content:字节	6.538
content:int	content:而	7.037
content:main	content:printf	7.307
content:name	content:struct	8.181
content:sizeof	content:字节	7.0
content:位	content:运算	6.538
content:值	content:函数	6.666
content:值	content:变量	7.631
content:值	content:指针	6.578

图 6　关系发现结果

(3)本体建立模块

实验中提取 103 个领域概念和 72 组概念关系。由于其中存在一些无效或错误的概念，所生成的关系也仅限于表达概念之间的关联，本体中其它关系需要人工加入，因此在本体建立步骤中，以提取的关联组合为基础，对课程本体的概念、关系做修正和调整。具体调整策略如下。

1)修正本体概念。主要包括两方面的工作：①去除领域无关的或是无意义的概念，如“而”、“登录”、“注册”等，它们虽然是高频词汇，但对于本体构建并没有实际意义；②修正由于分词造成的错误概念，如“include”错误地表示为“includ”，“整型”错误地表示为“整”和“型”等。

2)依据关联强度对已有的关系进行分析，将其区分为“概念”、“对象属性”、“数据属性”和“实例”4 大类。虽然这一过程仍然需要领域专家的参与，但通过对大量语料的统计分析，借助关联强度调整本体结构，可以有效地避免领域专家的主观思维，同时也有助于发现潜在的关联，突破了传统的以教材知识结构为依据的手工构建方法的局限性。

基于上述策略，对自动提取的课程本体进行人工调整，然后将调整后的本体用 OWL(Web Ontology Language)描述，建立课程本体。

4.2　技术评价

为证明课程本体自动构建技术的有效性和执行效率，从课程本体质量和建立效率两个方面设计评价实验，对技术作评价分析。

(1)本体质量评价

在课程本体质量评价中，采用绝对指标分析和指标对比分析两种方式。绝对指标分析中以标准术语集为基准，计算自动构建本体质量指标。指标对比分析中将自动构建本体质量指标与人工建立本体质量指标对比。评价指标围绕本体的概念和关系选取，具体包括：可译系数 α、规模系数 β 和关系覆盖度 rd。其中可译系数和规模系数是表征概念的指标，关系覆盖度表征关系。3 个指标的计算如式(3)－式(5)所示：

$$\alpha = N1/N \tag{3}$$

其中，$N1$ 是自动构建本体中概念和标准术语表中概念匹配的个数，N 是自动构建本体的概念总数。

$$\beta = N1/L \tag{4}$$

其中，$N1$ 同式(3)，L 是标准术语表的概念总数。

$$rd = R/N * \% \tag{5}$$

其中，R 是自动构建本体中的关系总数，N 是自动构建本体中的概念总数。

以《C 语言程序设计》课程为实验对象，在绝对指标分析实验中，将谭浩强教授主编的《C 程序设计》作为章节蓝本，在自动收集的语料库中随机选取 60 篇涵盖主要章节的语料文档，经筛选得到 191 个专业词汇，以此作为术语集。此术语集由于涵盖了课程的各个章节，且不受阈值的限制，因此可近似地作为标准术语集。实验中得到的自动构建本体质量指标如表 2 所列。

表 2　自动构建本体质量

概念数(N)	关系数(R)	标准概念数	匹配概念数(N1)	可译系数(α)	规模系数(β)	关系覆盖度(rd)
103	72	191	89	0.864	0.466	69.9%

在可译系数的计算中，由于 103 个本体概念的计算中包含了诸如“定义”、“指向”等这类不属于专业术语的属性，因此其实际的可译系数要高于 0.864，具有良好的专业相关性。相比可译系数，规模系数偏低主要是由于阈值的影响。实验中设置的阈值用于选择领域内的高频词汇，据此筛选出的词汇具有知识的代表性，可以间接反映出领域内的重难点知识。因此，得到的 0.466 的规模系数。说明构建的本体规模较小、知识粒度较大，但并不影响本体的领域完整性，且可以通过调整阈值的大小得到规模更大、知识粒度更小的课程本体。

在指标对比分析实验中，以 “C 语言”、“C 程序设计”和“本体”为关键词检索知网数据库，通过综合分析所得文献，选择刘光蓉在《“C 程序设计”课程内容本体构建》[4] 一文中手动搭建的《C 程序设计》课程本体为对比对象。指标对比分析的实验结果如表 3 所列。

表 3　自动与手动构建本体质量的对比

技术	概念数(N)	关系数(R)	匹配概念数(N1)	可译系数(α)	规模系数(β)	关系覆盖度(rd)
自动	103	72	89	0.864	0.466	69.9%
人工	183	130	92	0.503	0.182	71%

由实验数据可知，自动构建技术在概念提取方面的性能优于人工构建技术，可能因为自动构建技术中可以通过调整阈值，使概念规模更小，而命中率更高。而人工构建技术概念的命中率和规模依赖于创建者的主观经验。在关系覆盖度上，两种技术性能相当，这可能是因为在一个课程本体中，关系规模具有一定的稳定性。所以虽然自动构建技术提取的关系的绝对数量小于人工构建技术，但关系的相对数量与人工构建技术近似相等，这从另一个层面说明，实验中关系覆盖度的阈值选取具有合理性。

(2)本体建立效率评价

在课程本体构建效率评价中，以本体建立时间为评价指标。在利用自动构建技术建立本体的过程中，时间消耗主要

在于索引建立、概念提取和关系抽取3个步骤。因此,本体建立效率分析实验中,分别计算这3个步骤的时间消耗。

选取5门课程作为实验对象。设置爬虫爬取深度为3,爬取网页数为500,实验结果如表4所列。

表4 本体自动构建主要处理步骤所需的时间(s)

课程	索引建立	概念提取	关系抽取	总时间
教学系统设计	673	5.9	13.6	692.5
教育技术学	519	4.2	11.4	534.6
教育心理学	537	4.2	11.5	552.7
C程序设计	611	5.3	12.7	629
计算机组成原理	492	3.4	9.5	504.9

从数据可以看出,索引建立的时间消耗最多,关系抽取的时间消耗其次,概念提取的时间消耗最少。这是由于索引建立的过程包含了爬取资源的发现、网页信息的抓取、下载资源的存储、数据内容的更新等过程,其时间规模受爬取深度和爬取网页数量的影响。理论上,概念提取算法的时间复杂度为$O(n)$,其时间规模受建立的索引文件大小的影响,当建立的索引文件较大时,时间消耗也较多。关系抽取算法的时间复杂度为$O(n^2)$,其时间规模受筛选出的概念数量的影响,因此概念提取的运行结果是影响关系抽取时间的因素之一。根据以上分析,在课程本体构建的过程中,应合理设置爬取规则和阈值,避免索引和候选概念集的冗余造成的资源浪费和时间消耗。

从表4实验数据可知,本体自动构建技术平均在10min内可以完成一门课程的本体构建,这比人工建立本体的效率高很多。因此,本体自动构建技术与人工构建相比,在工作效率方面具有明显的优势。

结束语 课程本体作为课程知识的一种重要组织技术,在智能学习系统中有着广泛的应用。目前,课程本体构建大多依赖于人工完成,本体质量受到专家经验的影响,而且构建效率较低。课程本体自动构建技术是解决这些问题的有效手段。近5年,Web课程资源数量的迅速增加为课程本体构建提供了丰富的数据源,使研究课程本体自动构建技术成为可能。

基于网络爬虫、中文分词和关联规则挖掘等技术,提出了一种课程本体自动构建技术。该技术首先利用网络爬虫自动抓取与课程相关的网页资源;然后通过中文分词技术在这些网页中抽取词汇,并统计词汇出现的频度,以频度大于阈值的词汇作为课程本体的概念;接着采用关联规挖掘算法发现词汇之间的关联,并计算它们关联的强度,取强度大于设定值的关联作为课程本体中概念间的关系;最后,利用本体建模工具建立课程本体OWL描述,完成课程本体的表示与存储。以《C语言程序设计》为例,设计实验对该技术做质量评价和执行效率评价。实验结果表明,该技术能够建立与人工构建技术质量相当的课程本体。在技术执行效率评价实验中,选取5门课程,分别计算它们的课程本体建立时间。据实验数据可知,该技术平均10min内完成课程本体的建立过程,具有较高的执行效率。

当然,本研究还存在一些不足之处。1)还没有找到最优词汇频度阈值和关系强度阈值,这一点将影响自动构建的课程本体质量。这一问题需要结合最优化理论和实验做深入研究。2)还未对其它学科的专业课程验证本体的正确性,因此该技术是否能够推广到其它学科领域还有待证明。这也将成为进一步研究的另一个重要问题。此外,构建本体的目的是支持知识系统实现自动推理。所以,如何将本体自动构建技术与自动推理技术有效结合来构建人工智能系统也是一个值得探索的问题。

参 考 文 献

[1] Gómez-Pérez A, Manzano-Macho D. A survey of ontology learning methods and techniques[R]. OntoWeb: Ontology-based Information Exchange for Knowledge Management and Electronic Commerce, 2003

[2] 程晓. 面向半结构化文本的领域本体自动构建研究[D]. 哈尔滨:哈尔滨工业大学, 2009

[3] 吕爽. 基于叙词表的医学领域本体的构建研究[D]. 吉林:吉林大学, 2011

[4] Missikoff M, Navigli R, Velardi P. The Usable Ontology: An Environment for Building and Assessing a Domain Ontology [M]// The Semantic Web—ISWC 2002. 2002: 39-53

[5] 刘磊. 基于模板的SSE_CMM领域本体自动构建研究[D]. 广州:南华大学, 2011

[6] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// Proc. of the ACM SIGMOD Conference on Management of Data. 1993: 207-216

[7] Hearst M A. Automatic acquisition of Hyponyms from large text corpora[C]// Proceedings of the Fourteenth International Conference on Computational Linguistic. 1992

[8] Joachims T. A probabilistic analysis of the Rocchio Algorithm with TFIDF for text categorization[C]// Proceedings of the International Conference on Machine Learning(ICML'97). 1997

[9] Kietz J U, Maedche A, Volz R. A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet[C]// Aussenac-Gilles N, Biébow B, Szulman S, eds. Proceeding of EKAW'00 Workshop on Ontologies and Texts. Juan-Les-Pins, France, 2000

[10] Suryanto H, Compton P. Discovery of Ontologies from Knowledge Bases[C]// Proceedings of the First International Conference on Knowledge Capture. 2001

[11] Staab S, Schnurr H-P, Studer R, et al. Knowledge processes and ontologies [J]. IEEE Intelligent Systems, 2001, 16(1): 23-46

[12] 陆文豪. 基于关系数据库的专业领域语义词典构建研究[D]. 上海:复旦大学, 2009

[13] 杜晶. 本体知识库的完全化过程研究[D]. 北京:北京交通大学, 2010

[14] 王雪. 中文领域本体构建方法研究[D]. 武汉:华中科技大学, 2012

[15] 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔科夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87-94