一种适用于复合术语的本体概念学习方法

李江华[1,2] 时 鹏[1] 胡长军[3]
(北京科技大学国家材料服役安全科学中心 北京 100083)[1] (江西理工大学信息工程学院 赣州 341000)[2]
(北京科技大学计算机与通信工程学院 北京 100083)[3]

摘 要 术语的提取显然在本体概念学习中起着重要作用,由于汉语文本中词与词之间没有明显的界限,使得领域术语特别是复合术语的提取尤为困难。针对传统提取方法缺乏语义支持、计算量大、准确率低等不足,提出了一种适用于复合术语提取的本体概念学习方法。首先利用自然语言处理技术过滤掉与术语无关的成分,对语句进行自然切割,为领域术语提取提供完整的候选数据集,以保证候选领域复合术语不被误分。在此基础上,根据术语的领域统计和分布特征,利用术语频率和信息熵进行多策略的领域术语筛选,经同义术语识别与合并,获得领域概念集。经实验验证,提出的方法能够以较高的准确率从领域文本中提取出领域单词术语和复合术语。

关键词 术语提取,术语筛选,复合术语,本体概念学习
中图法分类号 TP391 **文献标识码** A

Ontology Concept Learning Method for Compound Terms

LI Jiang-hua[1,2] SHI Peng[1] HU Chang-jun[3]
(National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China)[1]
(School of Information and Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)[2]
(School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China)[3]

Abstract Term extraction plays an important role in ontology concept learning based on text. Because of no clear boundary among words in Chinese text, domain terms, especially compound terms, are difficult to be extracted. Traditional term extraction methods usually need large amount of calculation and lack of semantic supporting. A novel ontology concept learning method for compound terms was presented in this paper. At first, natural language processing technology is utilized to remove the irrelevant parts to get candidate terms. Sentences in the text are cut by punctuation marks and removed parts, so that the candidate compound terms can be reserved from wrong cutting. The candidate domain-specific terms are filtered by term frequency and information entropy with multi-strategy, according to the characteristics of distribution and statistics of terms. Then domain-specific concept set is obtained after the synonymous terms recog-nition. Experimental results show that the method can extract domain-specific word terms and compound terms with higher precision.

Keywords Term extraction, Term filtering, Compound terms, Ontology concept learning

1 引言

随着互联网技术的发展,信息呈现爆炸式增长。如何从中发现知识并进行形式化表示,从而实现人机之间以及机器之间的信息交互、共享与重用成为一项重要的研究课题。基于此,研究者把目光瞄准了本体(ontology)。本体是共享概念模型明确的形式化的规范说明[1],本体明确地描述了领域甚至更广范围内的概念以及概念之间的关系,使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义与公理[3]。本体为人机以及机器之间的交流提供了有效的途径,因此被广泛地应用于知识工程、信息检索、机器翻译、信息交换、软件工程以及自然语言处理、智能信息集成等领域。领域本体可以由领域专家通过手工方式构建,但手工方式构建费时费力、易出错,且维护困难。为此,研究人员提出本体学习的概念,即自动或半自动构建本体的一系列方法和技术,它通过利用各种数据源以自动或半自动方式新建或扩充来改编已有本体,由此构建一个新本体[2]。本体学习的主要任务包括领域概念、概念之间的关系以及公理自动或半自动获取,本体学习能够显著提高本体构建的效率,因此成为近年来的一个研究热点。

到稿日期:2012-07-17 返修日期:2012-10-17 本文受国家"十二五"科技支撑计划项目(2011BAK08B04),中央高校基本科研业务费专项资金资助项目(FRF-TP-12-162A),江西省教育厅科技项目(GJJ12345)资助。

李江华(1976-),男,博士生,副教授,主要研究方向为语义Web、数据工程与软件工程等,E-mail:jxsimil@hotmail.com;时 鹏(1977-),男,博士,副研究员,主要研究方向为语义Web、材料知识与数据工程;胡长军(1963-),男,博士,教授,主要研究方向为并行计算与并行编译技术、网格计算、云计算与服务计算、数据工程与软件工程。

在概念学习中，最重要的一步就是术语的获取。术语是专业领域中概念的语言指称，可以是单词，也可以复合词，即文本中的名词或基本名词短语，具有很强的领域性，能够表征领域的重要特征。复合术语是指由多个词组合而成的术语，这些词可能与领域相关，也可能与领域无关，但其组合结果所表达的语义构成了领域的重要概念。如："微观失效形貌"与"材料服役环境"都是复合术语，具有完整的语义，表征了材料服役安全研究领域的重要概念，其组合成份"微观/失效/形貌"与"材料/服役/环境"都不能表达与其相同的语义。

由于中文在词法、句法构造及语法上区别于西文，如词与词之间没有明显的界限、词语缺乏形态变化、词序严格等，使得基于西方语言的较为成熟的术语抽取和关系学习理论与技术并不完全适合于中文。由于汉语文本中词语之间无明显界限，因此中文分词技术是术语抽取的基础。中科院计算所的ICTCLAS中文分词系统是目前分词正确率最高的系统之一，作为对普通文本的分词，其可以达到最好的效果。但由于专业领域词汇的缺乏，在分词过程中，一些能够代表领域特征的重要专有复合术语常被错误地切分成多个单词，从而失去了其本身表征的领域含义。这样的切分失当对于术语的提取影响很大，使得学习效果较差。如上文提到的"微观失效形貌"和"材料服役环境"，在进行分词时分别被分成了"微观/失效/形貌"和"材料/服役/环境"，从而失缺了概念本身应有的语义。一些研究者把分词后的结果直接作为候选术语，忽略了分词过程误分的影响，从而失去了重要领域概念的抽取可能性。

本文以本体概念的学习为研究对象，结合自然语言处理和统计学技术，提出了一种适用于复合术语的本体概念学习方法，力图提高从领域文本中发现并抽取出领域术语，特别是复合术语的性能，从而提高领域本体概念学习的正确率。本文第2节综述了当前国内外的相关研究状况，分析了当前研究中存在的问题；第3节详细论述了本文提出的适用于复合术语的本体概念学习方法；第4节设计了相关的实验，对本文提出的学习方法进行了验证、比较和分析；最后总结了本文的工作并提出了下一步的研究设想。

2 相关研究

国外在本体学习研究方面起步较早，目前已经形成了较为成熟的理论与技术，一些典型的基于文本的本体学习工具也已经被开发出来。Amir Kabir 设计了基于波斯文的文本学习工具 Hasti[5]，该工具将自然语言处理和统计方法相结合，利用逻辑推理、模板驱动、语义分析等方法从文本中抽取概念及其关系，是为数不多的一个能够获取本体公理的工具。意大利 Rome 构建的基于英文的 OntoLearn 本体学习工具[6]，其核心是采用机器学习的方法从文本集中抽取领域术语，使用通用本体 WordNet 对抽取的术语进行解释，从而形成概念及其之间的关系。德国 Karlsruhe 开发了基于英文和德文文本的本体学习工具 TextToOnto[7]，该工具集成了多种算法，采用机器学习、统计、模式匹配等方法从文本中抽取概念及其关系。

近年来国内学者在基于中文的本体学习方面也做了大量研究。陈文亮等人提出利用 Bootstrapping 的机器学习技术[8]，从大规模无标注真实语料中使用种子概念和模式匹配的方法自动获取领域词汇。张锋等利用互信息(MI)方法实现中文术语抽取[9]。杜波等人采用统计分析与自然语言规则相结合的方法实现了术语抽取[10]。程勇等人研究开发了本体学习工具 OntoSphere[11]，该工具采用基于 HowNet 和统计学的学习方法从领域文档中学习专有术语。刘柏嵩完成了实验系统 GOLF[12]，该系统基于 TextToOnto 本体学习工具包，在系统中集成了术语相关频率(RTF)、词频逆文献频率(TFIDF)、信息熵(Entropy)以及 C 值/NC 值等统计方法进行领域概念学习实验。何婷婷等人提出了一种基于质子串分解的中文术语自动抽取方法[13]，其使用 C 值的方法获取候选词串，然后采用分解和统计的方法抽取领域术语。张春霞提出了一种使用主动词和语义角色来提取领域概念的学习方法[14]。于娟提出了一种基于结合词性分析与串频统计的词语提取方法[15]。

综上所述，基于文本的本体概念学习方法主要有基于语言学的方法、统计方法和二者相结合的混合方法。基于语言学的方法通常是采用自然语言处理技术，一般步骤有文档预处理、分词及词性标注、停用词处理；根据语言特点及语法结构来构造模板或规则进行短语识别等。其优点是不依赖于语料库，准确率高，计算量小，能较好地反映语义信息；缺点是分词准确性多依赖于词典，规则需要手工建立，很难保证完备性与一致性，针对不同的语言可移值性较差。基于统计的方法在对大量语料进行分析的基础上，根据领域词汇与普通词汇在语料库中不同的统计特征识别领域术语，常用的统计方法有：串频统计、互信息、TFIDF、RTF、Entropy 和 C 值/NC 值、领域相关度与一致度等。基于统计的方法适用于对大规模语料进行处理，不需要词典，不受语言限制。而其缺点是缺乏语义逻辑，计算量大，依赖于选取的语料库，准确率不高，低频术语常被过滤掉。混合方法可以兼顾二者的优点，既适用于对大规模语料库的处理，又能保持一定的语义逻辑，但仍然依赖于词典和规则，统计部分的语义支持和准确率有待于进一步提高。

上述方法虽然能够进行本体学习，但在候选复合术语串的发现问题上，有的研究者没有予以考虑，而有的研究者结合互信息或 C 值采用统计的方法去发现候选串，一方面缺乏语义支持，另一方面在统计的过程中采用词汇组合的方式，计算量大，目标性不强，漏统、误统率高。这些问题都会造成领域术语提取，特别是复合术语的提取准确率与召回率较低。本文提出的概念学习方法是基于语言学方法与统计方法相结合的混合方法，在进行算法设计时避免了上述问题的出现，能够以较高的准确率从领域文本中提取出领域术语，特别是领域复合术语。

3 适用于复合术语的本体概念学习方法

本文提出了一种适用于复合术语的本体概念学习方法，用于从文本中抽取领域术语，尤其是复合术语。该方法的基本思想是：为了提高领域复合术语抽取的准确率，在分词的过程中，一方面要保证获取文档中所有可能与领域术语有关的名词和名词串，另一方面要保证这些名词串不被错误地切分。该算法首先对文本进行中文分词和词性标注，但并不以分词的结果作为词语分割的标准，而是以标点符号分隔的句子为基本处理单位，依次过滤掉与术语提取无关的成分，对句子进

行自然切分，对切分结果形成的字串，使用统计和词性分析的方法提取其中的名词和基本名词短语。根据赵军等的研究结果[16]，基本名词短语的构成定义如下：

定义 1 基本名词短语(以下简称 baseNP)

baseNP→动词＋名词

baseNP→baseNP＋baseNP

baseNP→baseNP ＋名词

baseNP→限定性定语＋ baseNP

baseNP→限定性定语＋名词

其中，限定性定语→形容词|区别词|动词|名词|处所词|(数词＋量词)。

定义用表达式递归描述了基本名词短语的词性构成形式，如，基本名词短语可以由名词加名词构成，也可以由基本名词短语加基本名词短语构成等。该定义将作为候选术语抽取过程中基本名词短语的识别依据。

本文的学习算法自上而下分为 4 个层次：①候选术语抽取，对语料库进行分词及词性标注、无关项删除等处理，为领域术语筛选生成候选术语数据集；②领域术语筛选，经过①得到的候选词汇中含有大量的非领域词汇，领域术语筛选就是从中筛选出能够代表领域特征的词汇，产生候选领域术语集；③同义术语识别，术语是语法层面的处理单位，多个同义术语可以用一个概念来表示，如“电脑”和“计算机”表达的是同一语义，而概念是语义层面的处理单位，一个概念只表达一个语义，可以对应多个同义术语，因此需要识别合并多个同义术语为一个概念；④专家审核，为提高领域概念的准确性，还需要领域专家对机器学习生成的候选领域概念进行人工审核。本方法的具体处理流程如图 1 所示。

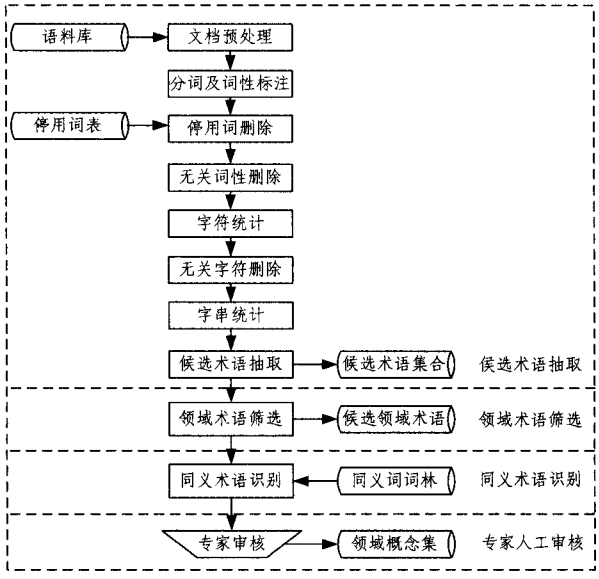

图 1 本体概念学习流程

3.1 候选术语抽取

候选术语抽取的目标是从语料库文本集中抽取出所有可能的名词和基本名词短语，抽取算法描述如算法 1 所示。

算法 1 候选术语抽取算法

Step1 文档预处理。把不同格式的待处理文档去除插图和表格等，将其转变为统一的纯文本格式。

Step2 中文分词及词性标注。本文所使用的中文分词软件是中科院计算所的 ICTCLAS2011 版。

Step3 删除停用词。停用词是指那些在各领域中出现频率都很高。并且与领域概念抽取无实际意义的字或词，如：“的”、“是”、“些”、“然而”、“因此”等一类词。在算法中可以对停用词表进行扩充，被删除的字或词使用空格代替。

Step4 根据标注的词性删除语句中的冠词、连词、介词、助词、代词和感叹词，因为这些词性不参与构成名词或基本名词短语[17]，与领域概念无关，被删除的字或词使用空格代替。

Step5 统计剩余文本中每个字符出现的频次，删除只出现一次的字符，使用空格代替，目的是为了提高语句的自然切分度。/* 在一个包含多个领域文档的语料中，只出现一次的字符尽管也有可能是领域术语，但这种可能性是非常小的。为了提高自然切分度，将其删除是合理的，即便不删除，在任何统计方法中，其都会因为词频太低被过滤掉。*/

Step6 删除单独出现的动词、副词、形容词，使用空格代替。所谓单独出现，是指这些词的两端不与任何其它汉字相连，这些词与名词短语无关。

Step7 对于剩余的被空格和标点符号切割开的字符串，统计每个串的长度与出现的频次，相同的串只保留一个，频次加 1，去除重复串；被长串包含的短串频次加 1。

Step8 按长度对所有的串进行升序排列。根据串长(即串所含的字数)分别做如下处理：

①一字串和二字串，如果该串是名词且频次大于指定的阈值，则入库为候选术语；否则，舍弃。

②三字串，如果是名词或构成名词短语且频次大于指定的阈值，则入库为候选术语；否则，该串的分词结果 a|bc 或 ab|c 交由①处理。

③四字串和五字串，如果该串是名词或名词短语且频次大于指定的阈值，则入库为候选术语；否则，根据该串在句中原始分词结果，交由①或者②处理。

④六字及六字以上串，如果该串是名词或名词短语且频次大于指定的阈值，则入库为候选术语；否则，根据名词短语的词性构成，提取其中的最长基本名词短语，对串进行二次切割，对每种可能的切割结果根据串长交由①或者②或者③处理。

由此完成了候选术语的抽取，由于在抽取的过程中首先去除了与术语抽取无关的各种成分，对语句进行了自然切割，因此不会对可能的复合术语造成误分。一方面采用统计方法结合名词短语的词性构成，根据重要的领域术语在本领域文本中出现频率必然高这一原则，保证了候选复合术语集的完整性；另一方面，在抽取候选术语时参照分词软件的分词结果，避免了个别不构成词汇的串因出现频率高而被误统为词汇，降低了错误率。

3.2 领域术语筛选

从上述步骤中抽取出的候选术语都是名词或名词短语，一部分是领域内术语，也有一部分可能不是领域内术语，因此需要对抽取的术语集进行过滤，以筛选领域相关术语。领域术语的过滤是根据领域术语在领域文档集与非领域文档集中不同的统计特征来实现的。术语的过滤应该遵循 3 个原则：

①领域术语在领域内文档集中出现的频率应大于在非领域文档集中出现的频率；

②领域术语由于其不同的领域特性，在不同领域中的分布是不均匀的；

③领域术语在领域内文档之间的分布是均匀的。

为此，在前人研究的基础上[18]，提出进行领域术语过滤的 3 个步骤：

为了便于说明，假设领域集 $D=(D_1, D_2, \cdots, D_n)$，其中 D_1 表示目标领域，其它为反例文档领域。

(1)首先过滤掉非领域术语,即那些在领域内相对于其它领域出现频率较低的候选术语。有部分候选术语可能同时是多个领域的术语,其在相关的其它领域中出现的频率可能高于本领域。为了减小这种候选术语被过滤掉的概率,同时为了避免领域语料规模对候选术语出现频率的影响,选择如下方法进行过滤:

①计算候选领域术语在各领域中的平均出现频率。

$$A(t|D_i)=tf_{t,i}/m\ (i=1,2,\cdots,n) \tag{1}$$

式中,$tf_{t,i}$表示术语 t 在领域 D_i 中出现的频率,m 表示领域 D_i 内文档数目。

②候选术语在领域内的平均出现频率应高于在其它领域内平均出现频率之和的均值,即按式(2)进行过滤,将不满足条件的滤掉。

$$A(t|D_1)>\sum_{i=2}^{n}A(t|D_i)/(n-1) \tag{2}$$

(2)术语在领域间的分布采用术语的熵值来计算,如式(3)所示。

$$GDC(t|D)=-\sum_{D_i\in D}P(t,D_i)*\log_2(P(t|D_i)) \tag{3}$$

$$P(t|D_i)\approx tf_{t,i}/\sum_{i=1}^{n}tf_{t,i} \tag{4}$$

式(3)的值越小,说明术语在领域间的分布越倾斜,其越有可能是某些领域的术语;相反,值越大,说明术语在领域间的分布越均匀,说明术语可能是领域间的一个通用词汇,应被过滤掉。

(3)设有领域 D_i 的文档集为$(d_1,d_2,\cdots,d_n)$,术语在特定领域文档中的分布如式(5)所示。

$$DC(t|D_i)=-\sum_{d_j\in D_i}P(t,d_j)*\log_2(P(t,d_j)) \tag{5}$$

$$P(t,d_j)\approx freq_{t,j}/tf_{t,i} \tag{6}$$

式中,$freq_{t,j}$表示术语 t 在第 j 个文档中出现的频率。式(5)的值越大,说明术语在该领域内文档中分布越均匀,其越有可能是该领域内术语;值越小,说明术语在领域内分布不均匀,其可能不是领域内术语。

通过第(1)步过滤,过滤掉了那些在考察领域内出现频率低于反例领域中出现频率的候选领域术语,这部分候选术语不是领域术语;通过第(2)步过滤,过滤掉那些在各领域出现频率都较高的通用词汇;通过第(3)步过滤,过滤掉那些在领域个别文档中出现频率较高的候选领域术语,因其分布的不均匀性,这部分术语也是领域无关的。之所以采用这样的 3 个步骤逐次过滤并且不使用加权值最后过滤,一方面是因为语料集中往往包含大量的词汇,逐次过滤将会逐次减少下一步的计算量,避免加权最后过滤造成的大量无意义计算;另一方面,第二次过滤中值要越小越好,第三次过滤中值要越大越好,可能存在一些通用词汇,在某些领域内出现频率不高,但分布都比较均匀,颠倒过滤顺序会造成漏滤,并影响术语的抽取。

3.3 同义术语识别

同义术语的识别可分为两种方法。第一种是基于统计的方法,在大规模的语料库中,选取一组特征词,统计词汇上下文一定大小窗口内各特征词出现的频率,形成词汇的特征向量,通过计算词汇向量间的夹角余弦作为词汇之间相似或同义度的衡量。第二种方法是查阅词典或知识库,常用的有 WordNet、同义词词林和知网(HowNet)。WordNet 是一个英文的知识库[12],包含了语义信息,它对不同的词性划分了不同的同义词集合,每个同义词集合都代表一个基本的语义概念,并且这些集合之间也由各种关系连接。同义词词林是一个中文同义词典[19],其存在格式类似于 WordNet,把词汇以层次关系组织在一棵或几棵树中,树中的每个结点就是一个概念。与 WordNet 和同义词词林不同,知网中的概念是用义原来描述的,义原的组织形式是一种树状的层次结构,义原之间也存在着上下位关系、同义关系、反义关系等多种关系[20]。本文通过查阅同义词词林进行同义术语识别和合并,不足之处在于,同义词词林只能识别普通的同义词,缺乏对特定专业领域复合术语的支持。但由于领域复合术语同义词并不多见,因此从汉语同义词识别的角度看,同义词词林是最佳的选择。

3.4 专家审核

本文所提出的概念学习方法采用的是自然语言处理与统计相结合的方法。在学习的过程中,一方面使用了分词软件和统计方法,由于分词的正确率达不到 100%,存在着误分,统计方法的使用也存在着误差;另一方面,由于领域专业词汇的缺乏,利用同义词词典并不能识别出所有的同义术语。因此,为了提高领域概念抽取的准确性,还需要领域专家人工参与,对学习算法产生的领域概念进行审定。

4 实验与分析

4.1 语料及评价标准

为了构建材料服役安全研究领域本体,同时为了验证文中提出的候选术语抽取和领域术语筛选算法,本文收集了 127 篇材料服役安全研究领域的文档(共 22 余万字)作为术语抽取的正例文档,以从复旦大学语料库中选取的一部分语料作为反例文档,其包括以下领域:艺术 130 篇、交通 130 篇、环境 130 篇、经济 130 篇、医药 130 篇。实验评估方法采用广泛使用的准确率(precision)、召回率(recall)、和 F 指数(F-measure)。准确率是指正确抽取词汇数与抽取词汇总数的百分比,召回率是指正确抽取词汇数与语料中所有词汇数的百分比,F 指数是指准确率与召回率的加权机体平均值,具体公式分别如下:

$$precision=correct_{extracted}/all_{extracted} \tag{7}$$

$$recall=correct_{extracted}/all_{corpus} \tag{8}$$

$$F\text{-}measure=\frac{2\times precision\times recall}{precision+recall} \tag{9}$$

式中,$correct_{extracted}$ 表示从文本中抽取的正确的术语数,$all_{extracted}$表示抽取出的全部术语数,all_{corpus}表示语料库包含的正确的术语总数。

4.2 候选词语抽取实验与分析

本实验的目的是为了验证本文第 3.1 节所提出的算法从领域文档中抽取候选名词和基本名词短语的性能,算法设计的目的是尽可能为领域术语筛选提供全面的候选名词和名词短语,尤其是名词短语。在 127 篇材料服役安全研究领域的文档集中,分别使用 ICTCLAS、基于互信息和 C 值的方法与本文算法在同一硬件平台上进行了抽取比较,其中使用 ICTCLAS 分词结果只统计其中的名词,抽取结果如表 1 所列。相关的实验环境包括:双核 IntelT6570 CPU 2.1GHz,2GB 内存,160GB 硬盘,Window XP 操作系统和 java 编程实现。

表1 候选术语提取结果比较

方法	抽取总词数	正确抽取数	准确率	召回率	F指数	耗时
ICTCLAS	5745	5204	0.9058	0.9311	0.9488	8′38
MI	5661	5374	0.9546	0.9615	0.9580	26′52
C-value	5691	5283	0.9283	0.9452	0.8892	28′35
本文算法	5629	5482	0.9739	0.9808	0.9774	20′45

从实验结果看，本文的算法无论是在准确率还是在召回率上都是最优的。ICTCLAS2011抽取出的词汇较多，且准确率只有90.58%。这是因为在实验中，由于领域专业词汇的缺乏，ICTCLAS把一些领域复合词进行了误分，如"金相组织"被分成了"金相/组织"两个词，"洛氏硬度"被分成了"洛氏/硬度"两个词等。互信息方法通过组合相邻两个或更多的词，通过计算共现频率判断其组合方式是否可能构成候选复合术语。由于这种方式必须考虑一个串中所有相邻词语的组合情况，但并不考虑组合在语义上是否成词，也不考虑是否构成名词或名词短语，只要共现频率达到阈值，即认为它是一个复合词，因此抽取出的总词数多于本文算法，但正确抽取数偏低，计算量较大，耗时较长。而C-value通过计算串及其构成子串的相关性，来判断串是否可能构成候选复合术语，对串进行初步过滤，但串的分解及子串的形成都存在着多种可能性，完全的分解及组合需要大量计算，对于相关性的判断通过设定概率阈值，缺乏语义根据，因此不能保证计算过程获取的串是有意义的词语，并且耗时较长。本文算法由于在设计时首先过滤掉了与术语提取无关的成分，通过对语句的自然切分，保证了潜在的候选复合术语不被误分，又通过语法分析和串频统计选取候选术语集，在一定程度上增加了语义支持，提高了候选串成为有意义的名词或名词短语的准确度，同时避免了使用组合方式判断是否成词带来的开销，计算量较小，准确率和召回率都高于其它方法。在本文算法的实验中，"应力腐蚀裂纹"、"屈服强度"等专业复合术语都被正确地抽取出来了。

4.3 术语过滤实验与分析

本实验的目的是为了验证本文第3.2节所提出的术语筛选算法的过滤性能，算法设计的目标是从第3.1节算法抽取出的大量候选术语中尽可能准确地筛选出领域术语。为了便于说明，使用本文算法同TFIDF算法和文献[20]提出的DR+DC方法进行比较($\alpha=0.9$，$\beta=0.3$)。在实验中，tf的过滤阈值取300，第(2)步过滤的阈值取1.0，第(3)步阈值取2.5，实验结果如表2所列。

表2 领域术语筛选结果比较

方法	筛选词汇数	正确筛选数	准确率	召回率	F指数
TFIDF	342	261	0.7642	0.9751	0.8569
DR+DC	285	243	0.8537	0.9075	0.8798
本文算法	272	253	0.9297	0.9435	0.9365

从实验结果看，TFIDF因为仅仅是从词频的角度进行过滤，所以召回率比较高而准确率较低；DR+DC基于传统的领域相关度与一致度，但没有考虑术语在不同领域间的分布特征；本文的算法既从词频的角度过滤，又充分考虑到了领域术语在领域内和不同领域间的分布特征，所以不管是从准确率、召回率还是F指数，其都是优于TFIDF算法和DR+DC方法。

结束语 针对领域复合术语提取准确率和效率低的问题，本文提出了一种适用于复合术语的本体概念学习方法，该方法结合了自然语言处理与统计方法的优点。在候选术语的提取上，根据自然语言的特点，逐步过滤掉与术语提取的无关项，形成对语句的自然切分，确保潜在的复合术语不被误分。同时结合词性分析与串频统计进行候选术语识别，避免了使用组合词汇共现频率判断是否构成复合术语而带来的计算量大、误差大、缺乏逻辑语义的缺陷，保证了领域候选术语的正确性和完整性，为领域术语的筛选提供了较为完整的候选数据集。在术语筛选过程中，根据术语的领域特性，以及术语在领域内和不同领域间的分布特征，提出了基于多策略的3次过滤方法。经过实验验证表明，本文提出的基于中文文本的本体概念学习方法优于传统的方法，能以较高的准确率和效率从文本中提取出领域概念，尤其是复合术语。下一步我们将在领域文本中概念间关系的自动获取上开展研究，以完善领域本体的构建。

参考文献

[1] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. University of Twente, Enschede, 1997

[2] Gomez P A, Macho M D. An over view of methods and tools for ontology learning from texts[J]. The Knowledge Engineering Review, 2004, 3(19): 187-212

[3] Maedche A. Ontology Learning for the Semantic Web [M]. Boston: Kluwer Academic Publishers, 2002

[4] Frantzi K T, Ananiadou S. The C-Value/ NC-Value Domain Independent Method for Multi-Word Term Extraction[J]. Journal of Natural Language Processing, 1999, 6(3): 145-179

[5] Shamsfard M, Barforoush A A. Learning ontologies from natural language texts[J]. Int'l Journal Human-Computer Studies, 2004, 60(1): 17-63

[6] Navigli R, Velardi P, Gangemi A. Ontology learning and its application to automated terminology translation[J]. IEEE Intelligent Systems, 2003, 18(1): 22-31

[7] Maedche A, Staab S. Discovering Conceptual Relations From Text[C]// Proc. European Conf. Artificial Intelligence (ECAI-00). 2000, 1: 321-325

[8] 陈文亮，朱靖波，姚天顺. 基于BootstrapPing的领域词汇自动获取[C]//第7届全国计算语言学联合学术会议论文集. 哈尔滨，2003: 67-72

[9] 张锋，许云，侯艳. 基于互信息的中文术语抽取系统[J]. 计算机应用研究，2005，22(5): 72-77

[10] 杜波，田怀凤，王立. 基于多策略的专业领域术语抽取器的设计[J]. 计算机工程，2005，31(14): 159-160

[11] 程勇. 基于本体的不确定性知识管理研究[D]. 北京：中国科学院计算研究所，2005

[12] 刘柏嵩. 基于Web的通用本体学习研究[D]. 杭州：浙江大学，2007

[13] 何婷婷，张勇. 基于质子串分解的中文术语自动抽取[J]. 计算机工程，2006，32(23): 188-190

[14] 张春霞. 领域文本知识获取方法研究及其在考古领域中的应用[D]. 北京：中国科学院计算研究所，2005

[15] 于娟，党延忠. 结合词性分析与串频统计的词语提取方法[J]. 系统工程理论与实践，2010，30(1): 105-111

[16] 赵军，黄昌宁. 汉语基本名词短语结构分析模型[J]. 计算机学报，1999，22(2): 141-146

[17] 董强，郝长伶，董振东. 基于《知网》的中文信息结构抽取[EB/OL]. http://www.keenage.com/html/c_index.html, 2010

[18] 刘桃，刘秉权，徐志明，等. 领域术语自动抽取及其在文本分类中的应用[J]. 电子学报，2007，35(2): 328-332

[19] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法 [J]. 吉林大学学报，2010，28(6): 602-608

[20] 董振东，董强. 知网导论[EB/OL]. http://www.keenage.com/html/c_index.html, 2010

[21] 张玉芳，杨芬，熊忠阳. 基于上下文的领域本体概念和关系的提取[J]. 计算机应用研究，2010，27(1): 74-76