文章编号：1003-0077(2018)03-0101-09

DRTE：面向基础教育的术语抽取方法

李思良，许 斌，杨玉基

（清华大学 计算机科学与技术系，北京 100084）

摘 要：术语抽取从非结构化文本中自动抽取专业术语。该工作在中文分词、信息抽取、知识库构建中发挥着重要的作用。当前术语抽取方法很大程度上依赖于词的统计信息，由于基础教育学科中术语具有极强的长尾特性，导致基于统计的术语抽取方法很难抽取出处于尾端的术语。该文结合基础教育的学科特点，提出了 DRTE：一种利用术语定义与术语关系挖掘，综合构词规则与边界检测的术语抽取方法。该文以初高中的数学课本为数据源进行术语抽取，实验结果表明我们的术语抽取方法 F1 值达到 82.7%，相比目前的方法提高了 40.8%，能够有效地在中文基础教育领域进行自动化的术语抽取。

关键词：术语抽取；术语定义；术语关系

中图分类号：TP391　　**文献标识码**：A

DRTE：A Term Extraction Method for K12 Education

LI Siliang, XU Bin, YANG Yuji

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Term extraction is an essential task where terms are extracted automatically from unstructured text based on a specific domain. Previous methods largely rely on terms' statistic information. However, terms in k12 education area have serious long-tail effect, which makes it hard to extract terms at the tail part in methods based on statistics. In this paper, we propose DRTE, a method which focus on extracting terms from their definitions and relations. Our method also utilizes term-formation rules and boundary detection strategies. Experiments on math textbooks for middle school and high school reveal 82.7% on F1 performance of our method, which significantly outperforms the current method by 40.8%.

Key words: term extraction; term definition; term relation

0 引言

术语作为在特定领域内表达专业概念的约定性符号，在中文分词、句法分析等自然语言领域发挥着重要的作用。在构建领域知识库的过程中，术语作为领域内知识的主要体现，在知识实例的扩充工作中有着重要的地位。从非结构化文本中手工进行术语标注耗费大量人力与时间，且会存在因标注遗漏而导致召回率降低的情况。因此自动术语抽取工作受到了越来越多研究者的重视。

目前的术语抽取方法主要包含两个步骤。第一步是通过对字符串的单元性计算来获取候选术语；第二步则通过术语性这一衡量指标来抽取出真正的术语。其中单元性用来刻画特定字符串组合的稳定性，术语性是用来描述一个语言单位在该领域内的相关程度[1]。术语抽取工作已经在多个领域中进行了尝试，例如数学[2]、生态学[3]、生物医学[4-5]、信息科学[4,6]和自然科学[7]，这些方法大都是基于统计的方法。但当我们为基础教育知识库构建进行术语抽取时，发现术语的统计特征和其他专业领域中的术语有较大的不同。以数学学科为例，术语“三角形”在初高中课本中共出现 1 779 次，而术语“切点圆”则仅仅出现 3 次。数学教材中仅有少部分重要术语被反复使用，这种长尾特性会造成低频词的遗漏。此外，一些基础性术语如“面”“线”也被广泛地使用在其他领域，

收稿日期：2017-06-06　**定稿日期**：2017-08-14

基金项目：国家科技部 863 课题(2015AA015401)

这种现象会导致通用性高的术语会因为逆向文件频率而被认为是领域无关的词语。

基础教育的相关书籍以教授知识为主,内容蕴含了大量术语的定义与术语关系的描述。我们结合基础教育资源的这种学科特性,提出了DRTE:以挖掘术语定义与术语关系为主,综合构词规则和边界检测的术语抽取方法。我们首先对书籍进行定义抽取,从定义中生成初始的术语候选。之后会进行数次迭代操作,每一轮迭代中,进行如下的操作:在全文和术语候选中寻找带有术语关系指示的内容并挖掘出新的术语候选;从术语候选中综合构词特点与边界检测的方法抽取出新的术语;最后将新发现的术语添加到分词的识别中,并开始下一次迭代。当不再有新术语发现时,停止迭代操作。

我们的实验针对基础教育的数学学科,选用了初高中数学课本的电子化书本作为数据源。我们的抽取方法的F1值达到达到82.7%,相比目前方法提高了40.8%。本文的创新点主要包括:(1)提出了一种利用术语定义与术语关系的非监督术语抽取方法:DRTE;(2)通过利用术语的定义与关系的背景信息,避免了基础教育中大量低频术语带来的术语遗漏现象;(3)针对因中文分词误差导致的长术语抽取困难现象,提出了迭代式的术语抽取方法。本文内容组织形式如下:第一部分介绍术语抽取的相关工作;第二部分介绍我们的术语抽取方法:DRTE;第三部分介绍我们的实验;第四部分展示实验结果与分析;第五部分给出结论。

1 相关工作概述

术语抽取关注于简单术语(仅由一个词构成的术语)和复合术语(由多个词复合的新术语)的抽取。目前的术语抽取方法可以分为三种:基于语法规则型、基于统计型以及基于机器学习型。

1.1 基于语法规则型

术语作为一个领域内独立存在的语言单位,其构词的结构应该是稳定且有规律的。基于这种假设,我们可以通过挖掘这种语言上的规律来进行术语抽取。例如,可以通过分析生物学词汇的构词方式来构建出一套通用的生物学术语命名规则[8]。另一方面,一些特殊的构词部件(如前缀和特定的缩写)也被用来进行术语的抽取[9]。除了构词规则之外,词汇在句子中的上下文信息也可以用来生成抽取规则[10]。这些基于语法规则的术语抽取方法普遍具有较高的准确率。但由于术语构词规则多变,该方法的召回率通常都不高。

1.2 基于统计型

与领域相关的文档通常会针对一个或几个术语展开描述,因而术语在这些文档中的分布具有一定的统计特性。利用术语的不同统计特征,可以对术语的术语性进行衡量。例如利用TF信息的方法[11]、基于TF-IDF的方法[12]。为了解决复合术语的识别问题,C-value方法[13]在原有的统计信息中加入了术语长度和嵌套术语的考量。结合中文的特点,一些如互信息[11]、改进C-value[14]的方法也相继被提出。基于统计的术语抽取方法对于领域的背景知识要求较低,具有较高的召回率。但在面对基础教育领域时,由于相关的文档通常以系统教授概念为主,术语的统计规律与其他领域有很大的区别,导致现有的统计量并不能很好地筛选出该领域下的术语。为了应对这种情况,LiTeWi方法[15]提出了利用外部Wikipedia资源,通过实体链接的办法来进行术语筛选。但该方法受限于外部资源的术语覆盖度与实体链接的准确程度,F1值仅为36.8%。

1.3 基于机器学习型

基于机器学习的术语抽取方法通常将术语抽取与术语分类结合在一起。这些方法利用训练数据基于机器学习的方法来学习术语抽取的特征[16]。Conrado提出的术语抽取方法中使用了八个术语的语言学特征(如词性、中心词),七个术语的统计特征(如TF-IDF值、词的长度)以及四个混合特征(如C-value)[3]。对于这类通常为有监督学习的方法,如何获取优质的训练数据是关键。此外,如何选择适合进行术语抽取工作的特征也是该类方法的难点之一。

实际的术语抽取工作通常不是单独使用上述三种方式的某一种,而是将它们选择性地组合在一起。例如为了利用术语的语言学特征与统计上的趋势,采用了规则与统计相结合的方法。但是,上述方法在直接应用到基础教育术语抽取时,还存在着低频词难以抽取的问题。

2 方法

本节介绍面向基础教育的术语抽取方法:DRTE。与基础教育相关的书籍以向学生讲授相关知识为主要目标,其内容包含大量的术语定义与对

术语关系的描述。为了利用好这些信息，我们提出了以术语定义与术语关系挖掘为主的术语抽取方法。该方法是一个迭代的过程，每一步根据已有的术语集从术语的定义和术语间的关系当中，综合构词规则和边界检测的方法发现新的术语，并更新术语集。DRTE 方法的流程如图 1 所示，包括如下四个关键环节：(1)文本预处理；(2)基于定义与关系的术语候选获取；(3)基于构词规则与边界检测的术语筛选；(4)术语集与分词结果更新。下面具体描述这四个环节。

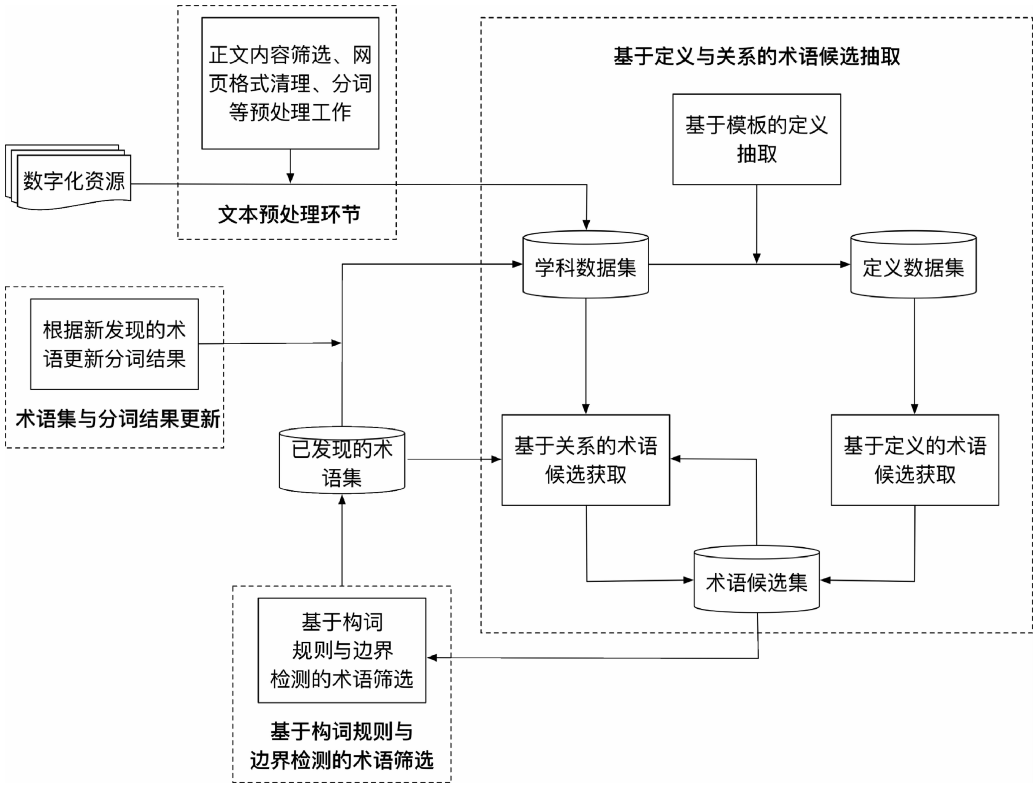

图 1　DRTE 方法的流程

2.1　文本预处理

我们的数据来源是基础教育课本的数字化 epub 资源。epub 资源的内容并非纯文本，而是以类似 HTML 网页的形式进行组织。故在利用这些数据之前，需要对其进行数据清洗。

我们首先筛选出书籍的正文部分(即不包括标题、前言、习题与单元总结的内容)，因为正文部分是这些书籍的知识主要来源。图片与表格中的内容也会从正文中删掉。为了避免公式对分词效果产生影响，我们用正则表达式过滤掉书中的数学符号与数学公式。之后去除了所有的网页标签，并根据句号、逗号、分号与问号对文本进行重新分段。最后，我们利用 ansj 分词工具[①]对文本进行中文分词，并计算得到每个词的词频。

2.2　基于定义与关系的术语候选获取

2.2.1　通过定义获取术语候选

我们首先从清理后的数据集中抽取定义。在我们的方法中，定义并不是获取术语候选的唯一途径，对于定义抽取的召回率要求不高，故采用模板来进行定义的抽取。表 1 展示了我们使用的模板：

表 1　用于定义抽取的模板

＜定义部分＞(叫|称)(做|为)＜被定义部分＞
＜定义部分＞是指＜被定义部分＞
＜被定义部分＞的定义(是|为)＜定义部分＞
称＜定义部分＞(做|为)＜被定义部分＞

通过模板匹配抽取出的定义会被分解为两个部分：被定义部分与定义部分。被定义部分揭示了该定义的描述对象，而定义部分则表示对描述对象进行定义的内容。

我们利用定义来获取术语候选基于如下两个假设：(1)课本中的定义都是用来讲授该学科知识的，故一定都是用来描述该学科中的术语的；(2)基础教

① https://github.com/NLPchina/ansj_seg

育学科中的术语应当呈现较强的自包含特性，即用来定义某一个术语的词语很可能本身也是术语。故对于一个定义，我们将其定义部分和非定义部分各作为一个术语候选。

我们以垂线的定义为例展示基于定义的术语候选获取。垂线的定义："取互相垂直的两条直线中的一条直线叫做另一条直线的垂线。"根据模板匹配，能确定被定义部分为"另一条直线的垂线"，定义部分为"互相垂直的两条直线中的一条直线"。根据上述的假设，这两个部分都能作为术语候选。

从上面的例子中可以看出定义部分和被定义部分的句子复杂程度是不同的。通常情况下，定义部分的句子更为复杂。此外，由于定义部分中经常混有公式，还会造成预处理后定义部分的结构并不完整。

针对上面的情况，尽管一条定义中能够产生两个术语候选，我们设置定义部分产生的术语候选为低置信度，被定义部分产生的术语候选为高置信度。在术语筛选的环节中，会根据不同的置信度等级采取不同的筛选策略。

此外，我们认为在定义部分与被定义部分产生的术语候选中，术语都应当处于靠右侧的部分，故它们均会被标记为右型候选(Rc)。左型候选(Lc)与右型候选(Rc)是用来指出术语更容易出现在术语候选的左侧部分还是右侧部分。在术语筛选阶段会根据术语候选方向的不同采取不同的策略分析。

2.2.2 通过关系获取术语候选

在该步骤中，我们根据已经获取到的术语集，结合术语之间的逻辑关系进行进一步的术语候选的获取。用于术语抽取的逻辑关系有三种：上下位关系、整体与部分关系及并列关系。

2.2.2.1 上下位关系

上下位关系指两个词之间体现出的语义包含关系。例如"正方形是一种特殊的长方形"中，"正方形"是下位词，"长方形"是上位词。我们通过模板"＜下位部分＞是＜上位部分＞位"来抽取上下位关系。如果匹配到的下位部分或上位部分中恰有一个部分是已发现的术语，则将其中不是术语的部分作为术语候选。例如在上例中，若"正方形"在已发现术语集中出现，则可以根据上面的规则，将"一种特殊的长方形"作为术语候选。

匹配到的下位部分会被标记为Rc，上位部分会被标记为Lc。由于能够反映上下位特征的句式并不一定都具有这种关系，例如"解三角形是一个重要的数学问题"中，匹配到的上位部分并不是一个真正的术语。故我们将上下位关系产生的术语候选设置为低置信度。

2.2.2.2 整体与部分关系

整体与部分关系通过"的"字短语来进行抽取。整体与部分关系中既存在"三角形的边"这样仅涉及术语的关系，也存在如"函数的难点"这样的有非术语参与的关系。在保证术语抽取准确度的前提下，为了更好地利用整体与部分关系进行术语抽取，我们根据抽取到关系的来源的不同，分别针对高置信度术语候选、低置信度术语候选与普通文本采取了不同的关系分析方法。

从高置信度术语候选中发现的整体与部分关系很有可能是在描述仅涉及术语的关系，故我们认为是最为可靠的，所以直接将"的"字短语中"的"左右两侧的内容均设置为高置信度的术语候选。"的"字左侧的内容标记为Rc，"的"字右侧的标记为Lc。

由于低置信度的术语候选通常句式会比较复杂，我们需要选择"的"字短语中句式较为简单、更可能存在术语的一部分作为术语候选。这里我们根据左右型候选来进行判断。若术语候选是Lc，则选择"的"字短语左侧部分作为术语候选并标记其为Rc，否则选用右侧部分并设置其为Lc。最后设置这个新发现的术语候选为低置信度。

从普通文本中发现的整体与部分关系往往处于句子的中段部分，关系的上下文较为复杂，很容易引入诸如"三角形的难点"这种类型的噪声结构。故对于这种类型的关系，我们采取了更严格的筛选措施。

由于从普通文本中获得的整体与部分关系中很可能并不存在术语，我们首先取出"的"字短语的左右两侧的词。这两个词中必须恰有一个是已经发现的术语。由于我们每次更新已经发现的术语集时都会重新更新一遍分词的结果，所以只要是已发现的术语，它一定会在分词时处理为一个词，而不会被切分成多个词语。故我们会将"的"字短语两侧中不是术语的词作为初选的术语候选。

之后，为了避免发生类似从"三角形的难点"中抽取出噪声术语候选"难点"的现象，我们会对上一步中得到的术语候选进一步进行候选可靠性检查。如果一个词是术语，那么与它有整体与部分关系的词中，术语应当占多数。基于这一假设，我们会检查所有有该术语候选参与的整体与部分关系，并根据已经发现的术语集统计其中非术语与术语的比值。若该比值大于指定的阈值 T_r，则判断该术语候选是应当剔除的。最终从普通文本中确定的术语候选将被设置为低置信度。

2.2.2.3 并列关系

我们通过模板："＜并列部分＞（＜并列部

分＞、)＊[和|或|与]＜并列部分＞等?”来识别并列关系。若并列部分中有一个为已发现的术语，则其他的并列部分也很有可能为术语。我们基于上面的假设将满足条件的并列部分作为术语候选。例如“棱锥与棱柱都是常见的几何体”中，若“棱柱”在已发现的术语集中，则将“棱锥”添加到术语候选中。由于并列关系中并列部分的句式结构通常较为简单，且一旦有一个并列部分为术语，其他并列部分为术语的可能性很高，故我们设置抽取出的术语候选为高置信度术语候选，并标记为 Rc。

在并列关系中，经常会出现术语中心词省略的情况。例如“锐角、直角和钝角三角形”中，中心词“三角形”就在前两个并列内容中被省略了。我们采取中心词检验的方法来处理这种情况。我们取出并列关系中最后一个并列部分，依次将其倒数最后一个字、倒数两个字，直至全并列部分作为中心词。例如上面的例子中，检验的中心词有“形”、“角形”、“三角形”、“角三角形”和“钝角三角形”。我们依次检查所有的中心词，将该中心词置于其余并列关系的尾部构成新的词语，并统计这些词语的出现次数之和。若和的最大值超过了给定的阈值 T_s，则认定该并列关系中出现了中心词省略现象。在上例中，当中心词为“三角形”时，“锐角三角形”和“直角三角形”的出现次数之和最高，故最终产生的术语候选为“锐角三角形”、“直角三角形”和“钝角三角形”。需要说明的是，我们不统计不带中心词的词语的出现次数之和，即只要认定了并列部分可以是“锐角三角形”，就不会考虑并列部分为“锐角”的情况。因为，虽然“锐角”和“锐角三角形”从语法上讲都可以看作是处于并列部分的术语，但在人的理解方式中，更倾向于用“锐角三角形”来进行理解。

2.3　基于构词规则与边界检测的术语筛选

基于术语的定义与关系抽取到的术语候选是从句式特征出发获取到的，并不能体现出术语作为词语本身的特点，因此还需要从构词规则与边界检测的角度对术语候选做进一步的筛选，以确定最终的术语。

2.3.1　构词规则

在平衡词性搭配规则的准确性与普适性的问题上，之前的研究工作主要采取了两种应对措施。一种方法是限制抽取的术语长度，如限制在 2～6 字之间。这种方法可以有效地减少可能的词性搭配情况，但会造成长术语的缺漏。另一种方法是适当宽松词性搭配规则的限制，但这种方法容易造成术语的误判。

我们称一个术语分词后的组成词语个数为该术语的元数。例如“三角形”是一元术语，而“直角三角形”则因为分词结果是“直角”和“三角形”而被定为二元术语。术语的元数会随着分词结果的变化而变化。我们在每一轮迭代中只考虑元数小于 4 的术语。在每一轮迭代结束后，会用已发现的术语更新分词结果。例如“单位正交基底”的初始分词结果是：“单位 正 交 基 底”，该术语是一个五元术语。但在第一次迭代结束之后，其分词结果为：“单位正交 基底”，是一个三元术语，故在第二次迭代中该术语候选就会被确认为术语。

我们参考的词性表是 ansj_seg 提供的词性表①。词性表包括 22 个大类，每个大类下有若干小类。后文提到的词性均指该词性对应的大类以及其包含的小类，在词性标注的过程中，我们发现很多领域术语的词性与分词工具标出的词性有很大区别。例如“边”通常会被标注为副词，但在领域中却应当作为名词。这种现象在基础教育领域的理科中尤为严重。因此在词性搭配规则的选取上，我们去除了常用的必须含有名词成分的限制，根据置信度的不同采用了更宽松的规则，如表 2 所示。

表 2　词性搭配规则

元数	置信度	词性搭配规则
一元	高	无限制
	低	非代词类 r、语气词类 y、助词类 u、连词类 c、叹词类 e、拟声词类 o、处所词类 s、状态词类 z、方位词类 f、时间词类 t 及英文词类 en
二元	高	无限制
	低	第二个词不为英文 两个词不是代词类 r、语气词类 y、助词类 u、连词类 c、叹词类 e、拟声词类 o、处所词类 s、状态词类 z、方位词类 f 、时间词类 t
三元	高	至少有一个词性为名词类 n、形容词类 a、动词类 v
	低	最后一个词性为名词类 n、形容词类 a、动词类 v 其余词不是代词类 r、语气词类 y、助词类 u、连词类 c、叹词类 e、拟声词类 o、处所词类 s、状态词类 z、方位词类 f 、时间词类 t

① https://github.com/NLPchina/ansj_seg/wiki/词性标注规范。

续表

元数	置信度	词性搭配规则
四元及以上	高	全部拒绝
	低	

低置信度的术语候选本身并不可靠，宽松的词性搭配规则容易降低术语抽取的准确性。故我们对置信度低的术语候选增加了术语命名规则。复杂的术语一般通过简单术语复合而成，故复杂术语的构词核心应当是一个术语。例如术语“离散型随机变量”的核心“变量”就是一个术语。通常情况下，术语的构词核心都在术语的后部，故我们会在已发现的术语集中寻找是否存在一个术语是该术语候选的后缀。如果不存在这样的术语，则在该轮迭代中不再考虑该术语候选。最后，我们会对低置信度的术语候选再进行一次词频的检测。我们会统计它们的出现次数，并选取出现次数高于给定阈值 T_c 的术语候选。

2.3.2 边界检测

学科的语言表达和词语搭配通常较为固定，这会导致一些领域无关的词语因为经常与特定术语搭配而被误认为是术语的一部分。例如“一条直线”就因为“一条”经常与直线搭配而被误认为是术语。与其结构完全一致的“一元方程”却是一个术语，这导致统计词首、词尾中特定字出现概率的方法失效。

我们选择手工建立边界词表来解决上述问题。边界词包括常见的副词（如“时”、“都”、“于”、“各”等）以及常用的代词和量词搭配（如“这个”、“一组”、“一对”、“一条”等）。

我们检查每一个术语候选。若该候选是 Rc，则从其分词结果的右侧起寻找到第一个出现在边界词表的词语，将该词右侧的部分作为新的术语候选，并删除原先的术语候选。例如，Rc 候选“一条 直线”中，从右侧起找到第一个边界词表中的词语是“一条”，则将其右侧的部分，也就是“直线”作为新的术语候选，并将原来的“一条直线”从术语候选集中删除。若候选是 Lc，则从左侧开始寻找，并挑选左侧部分作为新的术语候选。

通过上述的步骤，我们达成了两个目标：(1)对通过词性搭配检查的术语候选进行进一步分析，确定最终术语；(2)过滤掉四元及以上的术语候选中的一些边界信息，使其元数够降到四元以下。

2.4 术语集与分词结果更新

前三步结束后，这一轮迭代的术语发现工作已经结束。若术语集较上一轮相比没有发生变化，则终止迭代并输出最终的术语集。若术语集有更新，则利用这一轮中新发现的术语更新学科数据集和术语候选集中的分词结果。

我们会对分词结果中被分为几个词的术语进行修正，将其合并为一个词。新合并的词的词性根据合并前的最后一个词来判断。例如“异面直线”在合并前被分为“异面”和“直线”两个词。我们根据最后一个词“直线”来判断“异面直线”的词性。若最后一个词是名词类 n、形容词类 a 或动词类 v，则新词与其词性相同；否则新词的词性为名词类 n。如上例中，“直线”的词性是名词类 n，所以“异面直线”的词性与它相同，也是名词类 n。

更新分词结果之后，我们会重新计算所有词的词频，并进行下一轮的迭代。

3 实验

3.1 实验数据

我们选择基础教育的数学学科为研究对象，选择了人民教育出版社的初中数学课本 6 本，高中数学必修与理科选修课本 12 本，以及初高中教辅书 2 本，共计 20 本书的电子版。数字化的资源以 epub 格式（类似网页形式）组织。经过文本预处理后，共得到 7 万余个短句，共计 45 万余个词。

3.2 实验设置

对于从普通文本中发现的整体与部分关系，在术语候选的可靠性检查中，我们设置的阈值 T_r 为 0，即采取了最严格的术语检查。只有当与该术语候选之间有整体与部分关系的词均为术语时，我们才认为该候选是可靠的。这是由于在实验中，我们发现从普通文本中发现的整体与部分关系远没有从定义和术语候选中发现的整体与部分关系可靠。

在并列关系中，我们为术语中心词省略的处理过程设定的阈值 T_s 为“(并列内容数－1)×3”。例如在“锐角、直角和钝角三角形”中，我们会检查“锐角三角形”和“直角三角形”的出现次数之和是否大于 6。这里设置的阈值比较低，是由于并列内容的

句式较为简单，可靠性较高。将阈值设低一些能够有效地涵盖低频术语。

构词规则筛选中，对于低置信度术语候选的词频环节，需要设置一个好的阈值 T_c。为此，我们分别在不同的阈值下进行实验，得到术语抽取数量和 F1 值，如图 2 所示。可以发现 T_c 为 60 时 F1 值最大，故设置 T_c 为 60，即只有当该候选出现的总次数超过 60 时，我们才接受其为术语。

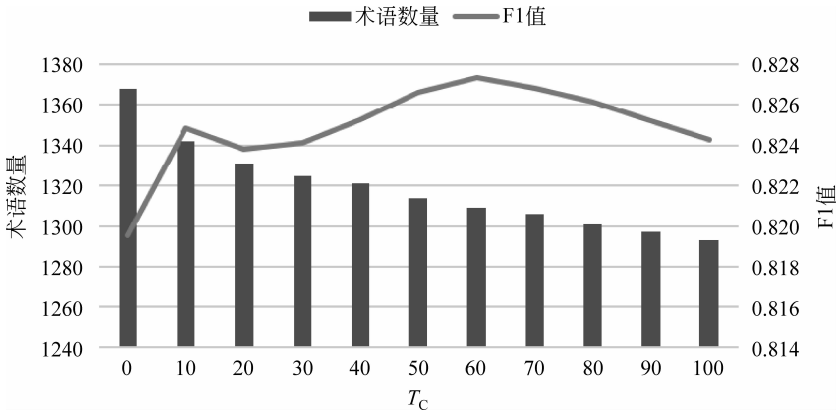

图 2　术语抽取数量和 F1 值随 T_c 的变化

3.3　评价方式

我们首先请基础教育数学老师对全部的课本进行一次标注，从中共标注出 862 个术语。之后请专家对由 DRTE 抽取出的术语进行审核，从中挑选出是数学基础教育领域需要涉及的术语。我们将人工标注的结果与 DRTE 抽取出的正确结果进行合并，作为书本中的术语全集。

由于基础教育领域中术语呈现显著的长尾特性，且如“点”“线”“面”这样的术语在很多领域中都有涉及。这导致目前大多数基于统计信息的算法都无法正常工作。我们选择了两个针对大量低频术语存在情况的术语抽取方法进行对比。LiTeWi 方法[15]通过与维基百科实体链接来提高低频术语的识别，基于信息熵和词频的方法[17]是一个针对中文术语的抽取方法。

4　实验结果与分析

表 3 展示了 DRTE 的实验效果。DRTE 共抽取出 1 186 个正确的术语，F1 值达到了 82.7%，效果相比之前的方法有了巨大的提升。根本原因在于我们改进了术语候选的获取方法。之前的方法为了照顾低频术语而引入了术语候选噪声，为此不得不采取了如与维基百科词条比对和信息熵的方法来提高术语的筛选能力。而我们的方法则从术语候选获取出发，通过定义来获取术语，并利用术语关系借助已发现的术语来识别未发现的术语，大大提高了术语候选的质量，进而提升了整个术语抽取的效果。

表 3　实验结果对比

	准确率/%	召回率/%	F1 值/%
LiTeWi	27.1	57.3	36.8
基于信息熵和词频	47.0	37.8	41.9
DRTE	**90.6**	**76.1**	**82.7**

为了展示出术语构词长度的分布情况，我们对抽取出的每个术语进行分词，统计构成该术语使用的词语数量，结果如表 4 所示。

表 4　术语构词长度分布情况

	术语个数	出现总次数	总词频/%
一元词	664	92 082	19.97
二元词	434	7 549	1.637
三元词	81	772	0.167 5
四元词	6	26	0.005 640
五元词	1	13	0.002 820

可以看出，术语多数是以 3 个以内的词构成的，最复杂的术语是由 5 个词构成的，术语的总词频超过了 1/5。

为了更好地说明为什么基于统计的方法不适用于基础教育中的术语抽取，我们对课本中的术语和所有词按照词频排序后绘制了词频的分布图，如

图3所示。尽管基于统计的术语抽取方法并不直接使用词频作为唯一的筛选,但该统计量在其他的复合统计量(如C-value、TF-IDF等)中有着重要体现,故我们选择词频进行分析。

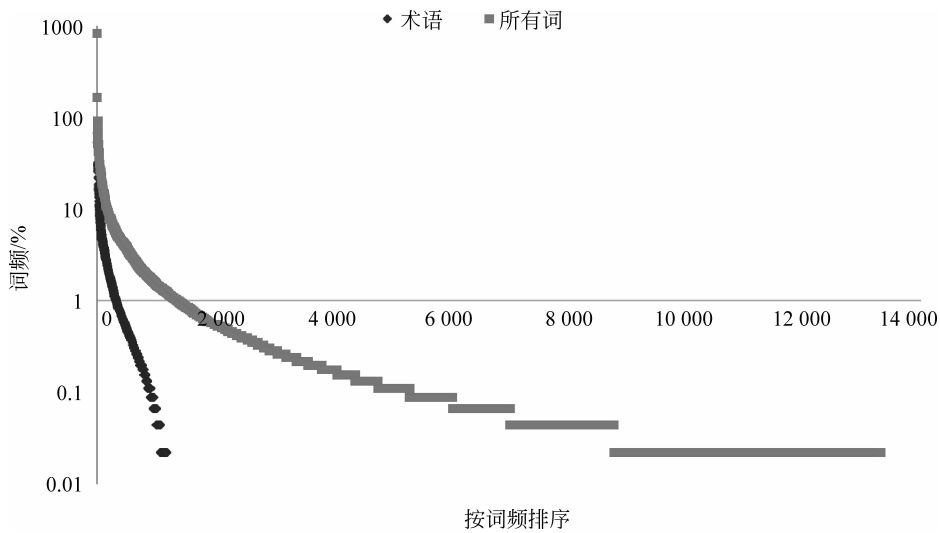

图3 词频分布图

可以看出术语词频在对数坐标轴下呈现近乎直线的分布,这说明术语词频有着指数级的下降速度,呈现明显的长尾效应。故基于统计的方法在提高方法的召回率时必须以低频术语的词频作为筛选标准,也就会导致大量非术语词汇的引入。

此外,可以看出术语词频的分布区间的下界与所有词词频分布区间一致,这说明处于尾端的术语的词频非常低。术语中词频排位在2/3的词语,在所有词的词频排位为3 500左右。而排位在3 500之后的词语本身也非常见词。TF-IDF统计量将很难区分这两类词,故基于统计的方法很难有效地筛选出术语候选。

我们从准确率和召回率两个方面来进行DRTE方法的误差分析。DRTE方法抽取错误的术语共有123个,经过分析可归纳为如下四种情况:

(1) 课本中存在领域无关的定义,如学习指数函数时,给出了"半衰期"的定义。这种情况仅出现9次,故我们"对课本中的定义绝大部分都是术语定义的假设"是比较可靠的。

(2) 固定搭配带来的误差。例如"函数的重点"中,"重点"一词的词频很高,而且与其构成整体部分关系的词均为术语。

(3) 边界检测的误判。我们发现基础教育领域中的一些术语具有多义现象,即在该领域中有特殊含义,在通常情况下却有不同的含义。例如"一次函数"中的"一次"指"最高项次数",而"一次独立重复试验"中的"一次"又有不同的含义。故边界检测无法判断这种类型的边界。

(4) 因分词造成的错误。一些句子在一开始就出现了无法纠正的分词错误。例如"其中大圆和小圆"就会被分词为"其 中大圆 和 小圆",导致误认为"中大圆"是一个术语。

在召回率方面,DRTE没有抽取出的术语可以分为三种情况:

(1) 一些术语的词频太低。如术语"周期数列"在课本中仅出现过一次。

(2) 一些术语虽然词频较高,但却未在定义与关系中多次出现。如"随机数"。

(3) 一些术语命名方式独特,与其他术语之间没有构词上的联系。这种类型的术语如果由多于3个词组成,则无法被识别,如"更相减损术"。

从整体的实验结果来看,我们的方法通过术语定义与术语关系抽取术语候选,充分利用已发现术语挖掘新的术语,能够解决大量低频术语存在的问题。实验结果证明了DRTE方法可以有效地应用于基础教育领域中的术语抽取工作。

5 总结

本文针对基础教育领域,提出了DRTE:一种利用术语定义与术语关系,综合构词规则与边界检测的术语抽取方法。为了解决基础教育领域中术语显著的长尾效应带来的对于低频术语召回困难的问题,我们结合基础教育以知识教授为主的特点,选择

从课本中术语的定义与关系来获取术语。我们分别介绍了从术语定义与术语关系中获取术语候选的方法，并阐述了基于构词规则和边界检测的筛选方法。随后我们介绍了实验的数据集与具体设置，并展示了最终的实验结果和相关分析。

实验结果显示：我们的方法在数据集上有着良好的表现，能够有效地进行面向基础教育的术语抽取工作。我们的方法对术语的词频依赖很低，能够有效地应对低频术语的情况。此外，我们的方法采取了迭代进行术语发现的策略，不断修正分词的结果，能够避免因分词误差而带来的术语遗漏。

参考文献

[1] Kageura K, Umino B. Methods of automatic term recognition[C]//Proceedings of the National Center for Science Information Systems. 1996: 1-22.

[2] Stoykova V, Petkova E. Automatic extraction of mathematical terms for precalculus[J]. Procedia Technology, 2012, 1(10): 464-468.

[3] Conrado M S, Pardo T A S, Rezende S O. Exploration of a rich feature set for automatic term extraction [C]//Proceedings of the Advances in Artificial Intelligence and Its Applications. Springer Berlin Heidelberg, 2013: 342-354.

[4] Lossio-Ventura J A, Jonquet C, Roche M, et al. Yet another ranking function for automatic multiword term extraction[J]. Lecture Notes in Computer Science, 2014, 8686(8686): 52-64.

[5] 孙水华，黄德根，牛萍. 中医针灸领域术语自动抽取研究[J]. 中文信息学报，2016，30(3)：118-124.

[6] 木合亚提·尼亚孜别克，古力沙吾利·塔里甫. 哈萨克语 IT 领域术语识别研究与实现[J]. 中文信息学报，2016，30(3)：68-73.

[7] Dobrov B V, Loukachevitch N V. Multiple evidence for term extraction in broad domains[C]//Proceedings of RANLP 2011. 2011: 710-715.

[8] Gaizauskas R, Demetriou G, Humphreys K. Term recognition and classification in biological science journal articles[C]//Proceddings of the Computional Terminology for Medical & Biological Applications Workshop of the 2 Nd International Conference on Nlp. 2000: 37-44.

[9] Krauthammer M, Nenadic G. Term identification in the biomedical literature[J]. Journal of Biomedical Informatics, 2004, 37(6): 512-526.

[10] Golik W, Bossy R, Ratkovic Z, et al. Improving term extraction with linguistic analysis in the biomedical domain[J]. Reseach in Computing Science. 2013, 23(4): 312-313.

[11] 张锋，许云，侯艳，等. 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)：72-73.

[12] 周浪，史树敏，冯冲，等. 基于多策略融合的中文术语抽取方法[J]. 情报学报，2010，29(3)：460-467.

[13] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International Journal on Digital Libraries, 2000, 3(2): 115-130.

[14] 胡阿沛，张静，刘俊丽. 基于改进 C-value 方法的中文术语抽取[J]. 现代图书情报技术，2013(2)：24-29.

[15] Conde A, Larra A M, Arruarte A, et al. Litewi: A combined term extraction and entity linking method for eliciting educational ontologies from textbooks [J]. Journal of the Association for Information Science & Technology, 2015, 67(2): 380-399.

[16] Zhang X, Song Y, Fang A C. Term recognition using conditional random fields [C]//Proceedings of the 2010 International Conference on. Natural Language Processing and Knowledge Engineering (NLP-KE), IEEE, 2010: 1-6.

[17] 李丽双，王意文，黄德根. 基于信息熵和词频分布变化的术语抽取研究[J]. 中文信息学报，2015，29(1)：82-87.

李思良(1991—)，硕士，主要研究领域为知识图谱、数据挖掘。
E-mail: lisiliang10@gmail.com

杨玉基(1994—)，硕士，主要研究领域为知识图谱、数据挖掘。
E-mail: yangyujiyyj@gmail.com

许斌(1973—)，博士，副教授，博士生导师，主要研究为知识图谱、数据挖掘。
E-mail: xubin@tsinghua.edu.cn