上海交通大学

硕士学位论文

专业领域术语抽取的研究

姓名：杜波

申请学位级别：硕士

专业：计算机软件与理论

指导教师：陆汝占

20050101

专业领域术语抽取的研究

摘　要

术语是通过语言或文字来表达或限定专业概念的约定性语言符号，是人类进步历程中知识语言的结晶。术语的抽取是术语库建立以及术语规范化的基础，是术语学与术语标准化工作的重要内容。由于现有语料数量多、动态性强，因此必须使术语抽取工作自动进行。

术语自动抽取领域采用的算法主要有两种，一是基于统计学的方法，这种方法认为，与普通词汇相比，术语拥有不同的统计特征（比如各组成成分之间较高的联系程度），依此可以鉴别出术语；一是基于规则的方法，这种方法认为术语表现出某些特别的形态语法结构或模式，其基本策略就是寻找和抽取结构符合某些特定模式的字符串。另外，还有一些算法将统计学方法和规则方法结合，取长补短。

在本领域内所作的工作大多针对印欧语系语言，对汉语进行的研究则相对较少。本文对汉语术语的抽取进行了一定的研究，设计了一个统计方法和规则方法相结合的汉语专业领域术语抽取算法，并具体实现。

在我们的实验系统 DSTES 中，共有四个核心模块：预处理模块，对系统的输入——生语料进行预处理，加入自定义标记；双字种子抽取模块，利用过滤词表对熟语料做进一步的操作，之后从中抽取符合统计标准的双字种子；双字种子扩展模块，对种子列表中的每一个双

字种子均执行扩充操作，得到候选多字术语项；后期处理模块，对多字术语候选项作最后的处理，过滤其中的非术语词汇。

系统首先利用统计模型从真实文本中抽取多字术语候选项，其中包含很多非术语项，因此必须考虑过滤操作。我们利用汉语语法规则和术语学原理，设计了一个多层过滤模型，其中使用的主要过滤技术有三个，这是我们的主要创新所在：符号和词类过滤，利用标点符号、特殊符号（如几何符号）及词类信息进行筛选；领域相关性和领域一致性过滤，考察候选项在特定领域及对比领域中的使用情况；模板匹配过滤，排除符合某些特定模版的候选项。另外，与以往系统不同的是，我们将过滤操作尽量提前，实验证明，这对于提高系统的效率是很有帮助的。

最终测试结果显示,本系统的抽取能力优于以前的方法,在开放测试情况下,MWU 的准确率达到 72.6％。

关键词 术语，自动抽取，规则方法，统计方法

Research on Domain-Specific Term Extraction

ABSTRACT

Term is a word or group of words having a particular meaning to express or restrict one professional concept in some specific domain, which is the fruit in human development progress. Extraction of terms lays foundation for term database establishment and term standardization and plays very important role in terminology. Because of the enormous capacity of today' s corpus as well as terms' dynamicity, it' s necessary to extract terms automatically.

In the field of automatic term recognition (ATR), there are mainly two types of algorithms: one is statistic-oriented and the other is linguistically -oriented. Statistic-oriented methods are based on the principle that terms have different statistical features, such as high relationship between segments of them, from common words, which can distinguish them. Linguistically –oriented methods are based on the principle that terms exhibit some special syntactical structures or models and their basic policy is to search and extract strings that match these specific models. There are still some algorithms combining statistical and linguistic methods. In another word, they are hybrid methods.

Most of the work in this field focuses on Indo-European languages, such as English and French, and the research on Chinese is much less. In

this thesis, we pay attention to Chinese term extraction, design a hybrid algorithm and finally implement it.

In our experimental system DSTES, there are four modules: pre-processing module, which preprocesses system' s input and inserts some system-specific symbols; bi-word candidates extraction module, which processes corpus by the help of filter-word list and extracts bi-word candidates; candidate extension module, which extends every bi-word candidate and gets multi-word term candidates; post-processing module, which filters non-term multi-word term candidates.

DSTES uses linguistically-oriented methods to extract multi-word term candidates with non-term ones, so it' s vital for our system to do some filter operations. According to Chinese grammar and principles from terminology, we design a multi-layer filter module that has three main technologies which are our main innovations: filter by symbols and word classes, filter by domain relevance and domain consensus and filter by templates. At the same time, we bring filter operations as forward as possible that is different with previous systems and great helpful for system efficiency according to our experiments.

According to final experimental data, DSTES is better than prior systems. On an open corpus, its precision reaches 72.6%.

Keyword term, automatic extraction, statistic-oriented method, linguistically-oriented method

第一章 概　述

当今社会是一个信息社会，科技飞速发展，网络应用日益深入广泛，它们极大地改变着人类的生存环境，各种社会观念、生产方式、社会分工发生了巨大的变化，我们正处于一个前所未有的“信息爆炸”时代。在这些海量信息中进行信息提取和知识挖掘，无疑具有至关重要的意义，而术语则在这个过程中扮演着重要角色。

术语是专业领域中概念的语言表示[1]，也可定义为“通过语言或文字来表达或限定专业概念的约定性语言符号”[2]。《中国大百科全书》中指出，术语是“各门学科中的专门用语。术语可以是词，也可以是词组，用来正确标记生产技术、科学艺术、社会生活等各个专门领域中的事务、现象、特性、关系和过程”。术语是科学研究的成果，是人类进步历程中知识语言的结晶，本文所作的研究工作就是围绕术语来进行的。

本章的内容是这样安排的：首先说明术语自动抽取技术的意义，然后介绍国内外术语自动抽取技术的研究现状，最后提出本文的研究目标，并简要叙述我们所做的工作。

1.1 术语自动抽取技术的意义

术语的抽取是术语库建立以及术语规范化的基础，是术语学与术语标准化工作的重要内容。如果完全由人工来进行术语的抽取工作，其弊端是显而易见的：

（1） 现有的语料浩如烟海，人工进行代价巨大，进展缓慢；

（2） 社会日新月异，术语动态发展：旧有术语逐渐消亡，新术语不断涌现，流动性强。

为解决上述问题，必须求助于拥有强大计算能力的计算机，求助于计算语言学。

计算语言学(Computational Linguistics)是在社会需求和技术进步的双重推动之下，由历史悠久的语言学和新兴的计算机科学相结合而产生的一门交叉学科。它伴随着计算机的发展而发展，距今已有近 50 年的历史。借助计算机强大的计算能力，同时结合语言学研究的成果，使术语抽取工作自动进行，对于术语学的研究来说，具有重要意义。

因为简洁一致的术语库是建立知识库的基础和捷径，是许多应用的出发点，所以术语自动抽取技术的应用前景非常广阔，涉及到机器翻译、自动索引、建立

词法知识库、信息检索等各个方面。下面列举了它在术语库的建立及术语规范化、生语料切分、机器翻译、信息检索、本体建立等五个方面的应用：

（1） 术语库的建立及术语规范化：随着社会的发展进步，旧概念不断消亡，新概念大量涌现；另外，在与外来文化交流的过程中，必须同时吸收外来术语。面对术语的急速增长和高速传播，为了避免交流与使用上的混乱，必须做好术语库的建立及术语规范化工作。借助术语自动抽取技术，我们可以方便而快捷的获取这些新概念和外来术语，建立术语库，并为随后的术语规范化提供基础；

（2） 生语料切分：由于中文文本是按字连写的，词与词之间不像英文那样存在间隙，要进行词性标注等后续操作，就必须首先解决词语的切分问题。自动分词就是将连续的、缺乏词语之间的分隔标志的生语料，自动切分后生成以词语为单位的熟语料。在切分的过程中，难点之一就是未登录词的识别。运用术语自动抽取技术，可以帮助识别未登录词，并在此基础上提高切分操作的效率和准确性；

（3） 机器翻译：简称 MT，是利用计算机进行的自然语言之间的翻译。在机器翻译的过程中，首先要对源语言语句进行分析，建立语法树。此时一个拥有大量词汇的词典是至关重要的，否则会导致歧义过多，影响翻译效果。术语自动抽取技术可以帮助自动建立词典，从而提高翻译结果的正确率与翻译过程的效率。参考文献[3]中实现了一个这样的试验系统，该系统先利用术语自动抽取技术获取一个领域字典，然后在该字典的辅助下进行该领域语言的翻译。另外，还出现了一些集成了术语自动抽取功能的商业翻译软件，如 TRADOS 公司的 TRADOS ExtraTerm 系统；

（4） 信息检索：信息检索（Information Retrieval），更确切的说是文档检索，就是通过将用户的需求与文档相匹配的方式来向用户提供相关文档（或相关信息）的过程。通常，文档和用户需求之间的匹配不是直接进行的，而是通过它们的代理（Surrogate）来匹配的[4]。因此，如何获取这些代理，就成了信息检索的关键问题。术语自动抽取技术无疑提供了一种比较理想的代理获取方式；

（5） 本体建立：本体（Ontology）最早是一个哲学的范畴，后来随着人工智能的发展，被人工智能界给予了新的定义。Ontology 的目标是捕获相关的领域的知识，提供对该领域知识的共同理解，确定该领域内共同认可的词汇，并从不同层次的形式化模式上给出这些词汇（术语）和词汇之间相互关系的明确定义。由此可见，领域词汇

的获取，对于本体的建立具有至关重要的意义。此时，术语自动抽取技术可以为本体的建立提供良好的素材。

1.2 国内外术语自动抽取研究的现状

针对术语自动抽取领域的特点，继 H.P.Luhn[5]之后，世界各国的研究人员做了许多艰苦的工作以及有益的探索与大胆的尝试，提出了一些各有特色的方法。根据计算语言学理论，这些方法大致可以分为如下三类：

（1） 基于规则的方法（Linguistically-Oriented Methods）。这类方法利用规则在语料中进行匹配，将符合既定规则的多字单元（Multi Word Unit, MWU）作为术语输出。它的优点是：

- ✧ 简洁直观、表达能力强；
- ✧ 可应用专家知识；
- ✧ 在先验知识与文本匹配的情况下，准确率高。

它的缺点是：

A. 适应性不强，因为不同领域的先验知识不同；

B. 无法识别未知词汇。

另外，由于相对于一般领域内的通用词汇而言，专业词汇往往是未登录词语，且针对性较强，在它的识别过程中有以下几个难点[6]：

A. 构词无规律。专业术语的构成不像人名、地名有一定的构词规律。它的构成方式多样，有些是由单字词或语素字组成；用字比较分散，有些是普通字，有些是生僻字；专业术语的长度也没有一定的限制；

B. 缺乏启发信息。人名、地名的识别有一定资源《中国人名用字库》、《中国地名用字库》可以借鉴。并且同中国人名相比，缺乏像姓氏一类的启发信息；

C. 专业术语指示词出现情况多样化。在真实文本中，一些介词、动词之类的指示词（防治、危害）经常同专业术语一起出现，对专业词汇识别能起标志作用，但这类词在文本中并不总是与专业词汇同时出现。如：“危害广大人民群众的利益”、“防治策略有很多”；

D. 专业特征词出现情况复杂。专业术语经常伴随着一些专业特征词出现，如：病、虫、蛾等等。但是文本中出现的专业特征词，并不都表示真正的专业术语。如：“重病”。

由上可见，种种现象使专业术语的识别变得复杂。因此，纯粹基于规则方法的算法相对来说不是很多。

参考文献[7]中认为术语表示非歧义的固定概念，并且具有特定的语法形式。在这篇文章中，他介绍了一个抽取法语术语的系统 LEXTER。该系统中针对的处理对象是法语文本，主要利用两个阶段来识别“术语单位”（Terminological Unit）：analysis 和 parsing，即在 analysis 阶段 利用“边界标志”（Frontier Marker）从文本中提取最大长度的名词短语，然后在 parsing 阶段从前述名词短语中提取可能的术语单位，最后由人工进行确认。

参考文献[8]中认为术语具有两个特征：①术语在使用时不会有形态上的变化，因此在语料中是以同一种形式出现的；②术语的语法结构局限于某种形式的名词短语。作者据此从标注后的文本中抽取出具有特定形式的、并且出现频率超过一定阀值的候选术语项；

（2） 基于统计的方法(Statistic-oriented Methods)。这类方法利用各种统计模型从概率意义上来衡量多字单元 MWU 是否为术语，其优点是易于实现；较少需要人的干预；适应性强，可用于不同领域；可识别未知词汇。缺点是不够简洁直观，对语料的依赖性很强：运行结果严重依赖于语料；必须有充足的语料才能获取较为理想的结果；准确率不高，因为许多概率意义上关联的词汇都不是术语；无法识别低频术语；由于必须进行大量的计算，很容易带来运行效率的问题。就整体而言，在术语自动抽取领域中已经提出的算法大多是针对技术术语及名词短语的，并且主要基于统计学方法。

参考文献[9]介绍了他们的 INDEX 系统。该系统抽取重复的单词序列，并根据一个经验公式赋以权重；

参考文献[10]提出了从名词短语中抽取“多字基本文本单元”（Multi-word Basic Text Unit, BTU，与术语学家理解的术语大致相当)的方法。他认为，在每个 BTU 出现时，都必须保持其各个组成部分的结构依赖性，而这种结构依赖性的衡量依据就是各种统计量；

参考文献[11]使用了两个统计标准：一个是扩展后的互信息（Mutual Information），可处理三个以上单词的情况；另一个是成本标准（Cost Criteria）[12]；

（3） 统计与规则相结合的方法(Hybrid Methods)。这类方法结合了上述两种方法的长处，近年来得到了较大的发展。其实严格来说，上面提到的对很多方法的分类并不是太严格：它们都不是纯粹的基于规则的方法或者基于统计的方法，都或多或少的借鉴了对方的思想和做法。我们做上述归类，只是说明它们的侧重点不同罢了。很多研究人员对这两种方法的具体结合方式作了积极的探索，以期达到系统整体的最佳效果。

参考文献[13]首先利用简单语法规则提取出来名词短语，然后应用一种称为C-value的标准来判断术语。

参考文献[14]中介绍了一个TRUCKS（Term Recognition Using Combined Knowledge Sources）方法。这种方法结合了规则信息和统计信息，用于判断上下文与候选术语的相关性，从而帮助提高识别术语时使用的传统统计方法的性能。

上面介绍的主要是国外研究者在本领域内所作的工作，国内对专业术语自动抽取、特别是针对中文专业术语进行抽取的研究工作则相对较少，下面作一概要介绍。

参考文献[6]中首先采用了统计的方法获取农业病虫害词汇的词性搭配规则、语义类分布规则，并进一步利用这些规则在大规模预料中采用并列同现、模式匹配、特征词匹配等策略获取病虫害词汇，建立特定专业领域（主要为农业病虫害领域）词汇词典。

参考文献[15]中提出了一种 Bootstrapping 的思想（“步步为营”），即首先以少量的领域核心词为“种子”，然后根据候选词与这些“种子”的亲疏关系确定结果术语，并选取新的“种子”加入种子集合，之后再据此从候选词中挑选结果术语和“种子”，这样就实现了滚动式的发展。

另外，国外学者也在汉语术语的自动抽取方面做了一些有益的尝试。参考文献[16]中设计了一个中文术语自动识别系统，它主要利用汉语复合词和它的各组成部分之间的关系来进行识别。

1.3 研究目标及我们的工作

经过各国研究人员近 50 年的辛勤工作，术语自动抽取领域的研究取得了很大的进展。根据本领域的特点，提出了许多各具特点的算法思想。这些算法思想在统计模型、语法规则等方面均有不同的设计思路和侧重点。另外，在处理对象上也有不同，有些算法需要经过标注的熟语料，另外一些算法只需要未经处理或只经过简单处理的生语料。在这些算法的基础上，实现了一些自动术语抽取器[17][18][19]。

如前所述，目前的术语抽取算法大多基于统计方法，这样做的好处是简单易行。但总体来看，现有的成果与人们的期望和需求还有一定的距离。由于自然语言本身的复杂性，现有的许多系统还没达到实用的地步。我们挑选了参考文献[20]中提到的统计模型，进行编程实现并用足球及金融领域的语料进行实验，在对实验数据进行分析后，发现如下主要问题：

1. 效率偏低，不论从时间效率还是空间效率来讲，均不够理想。经分析，主要是因为中间结果过于庞大；

2. 抽取结果不够理想，正确率不高。因为很多非术语词组，如“I am”、“你是”、“有一天”等，从概率意义上来说，与术语没有任何区别，因此无法区分。

另外，一个更为严重的问题是：有相当一部分结果，尽管是具有意义的词组，如“有一天”、“上海”等，但是它们并不是属于本领域的术语，对某些应用，如生语料切分有一定的意义，但是对于其他的应用，如专业术语库的建立、信息检索、本体的建立等方面没有任何意义。在通过这些算法得出结果后，除了必须剔出那些不成词的候选项以外，还必须排除这些成词的“无用项”。当抽取结果比较庞大时，这显然是违背“自动抽取”的精神的。

为达到实用的最终目标，针对上述问题，本课题从统计学方法和规则方法两方面同时入手：在对比现有的各种基于统计学的算法之后，我们决定采用参考文献[20]中使用的统计模型进行建模，并且根据实验结果对其进行了改进；同时，充分利用汉语语法规则和词法规则设计相应的过滤算法，集中精力于中间结果的控制，设计了一个统计方法和规则方法相结合的专业领域术语抽取算法，并具体实现。在获得详细的实验数据之后，对其进行分析，针对所得的结论对未来的工作做出展望，明确今后的研究方向。

本文的其余几章是这样组织的：第二章介绍基于多策略的专业领域术语抽取系统（Domain-Specific Term Extraction System, DSTES）的总体结构和各模块功能。三、四两章具体讨论本系统所用的方法。其中第三章讨论本系统中用到的多层过滤算法，第四章讨论本系统所采用的统计模型。第五章介绍测试结果并展望下一步的工作。

第二章　专业领域术语抽取系统（DSTES）

本章对我们实现的基于多策略的专业领域术语抽取系统（DSTES）进行总体介绍。其中第一节介绍了汉语术语的特征及其在语料中的分布规律；第二节讨论系统设计中要考虑的语言学相关问题和技术实现上的问题；第三节给出系统的总体结构图，然后对各个模块做详细介绍，作为以后各章讨论的基础；第四节简要介绍系统中使用的数据结构。

2.1　汉语术语的特征及其在语料中的分布规律

DSTES 系统的处理对象为未经处理的大规模特定领域生语料，主要针对汉语，也可以方便的迁移到其他语种。不同的语言有其共同的性质，也有很大的差异。因此在设计汉语自动处理系统时，必须充分考虑汉语的特性。要实现对汉语文本中术语的自动抽取，就必须研究汉语术语的特点及其在汉语文本中的分布情况。参考文献[21]指出，在专业领域文献中术语的分布主要有三种情况：

（1） 术语处于特殊的位置，比如关键词和注释中的术语等；

（2） 在新出现的或者作者认为比较新、比较难懂的术语后加上注释，并把注释用括号括起来，有的作者在自己的文章中第一次使用某个术语的时候，还要对术语进行解释；

（3） 术语无任何标记。

对上述各类分布的术语进行自动抽取的难度也不相同，第一类术语有明显的前界和后界，比较容易提取；第二类术语的右边界已经明确，需要确定该术语的左边界；第三类术语自动抽取的难度最大，因为提取这类术语不仅要确定前界和后界，还要判断这个语言片段是术语还是一般新词。

2.2　系统设计中的问题

2.2.1 理论问题

在设计特定领域术语抽取系统之前，我们首先对相关的概念问题作一些初步的考察。这其中，最为关键的一个问题就是：到底什么是术语？

如前所述，术语是“专业领域中概念的语言表示”，或者说是“通过语言或文字来表达或限定专业概念的约定性语言符号”。这个定义表明，术语实际上

描述的是专业领域范围内的“概念”、“指称实体”、“词语”三者的关系，它是一种“约定性语言符号”。

尽管定义明确，但是这些定义明显缺乏操作性。在实际操作时，由于学术背景、知识构成不同，专家们仍然会有意见相左之处。争议不仅存在于具体术语的定名及定义，更存在于学科发展的导向——它影响了术语收录的范围及定名、定义。这就容易导致混乱，比如计算机技术中的HTML，有人称之为“超文本标记语言”，有人称之为“超文本标志语言”。因此，术语由最初的产生、流行，直到最终被接受，本领域专家对其进行的规范化操作是必不可少的一个环节，由此产生了术语学，它用规范化的理论、原则和方法来指导术语规范化工作。一般认为，术语学作为一门学科，是奥地利术语学博士欧根.于斯特（Eugen Wuister）教授提出来的，他也是术语学中维也纳学派的创始人[22]。

我们在设计系统时，力争避开术语学方面的争论，而是从计算机处理的角度出发，将系统的设计目标确定为：从真实文本中尽可能抽取准确、全面的“术语候选项”，为术语规范工作提供良好的素材。与此同时，利用术语学方面的一般原则，如领域相关性、领域一致性等，设计了相关算法。但是这种考虑还只是停留在直观的阶段，缺乏术语学理论的指导，因此还有很大的提高空间，这也是我们今后要努力的一个方向。

2.2.2 技术实现问题

前面已经介绍过了，在设计特定领域术语抽取程序时，有三种方法可供选择：基于规则的方法、基于统计的方法、统计与规则相结合的方法。

基于规则的方法认为：术语表现出某些特别的形态语法结构或模式[23]。这类方法的基本策略就是寻找和抽取结构符合某些特定模式的字符串。由于这些模式在大多数情况下是与具体语言相关的，因此，基于规则方法要求针对具体语言作相应的处理。例如上面提到的参考文献[8]中的算法，需要先对语料进行标注，然后再利用单词的词性信息来进行语法模式匹配。

印欧语种在词汇、语法、语用、语境诸层面上有明显的界面区别，相互之间又有对应关系。这种情况可称为明显的分层性，简称为“面结构”。但是汉语则不同，各层面之间很难划分经纬，词法与句法之间没有明显的界限。另外，汉语也缺乏时态、语态词形变化等信息。这就使得分析并使用汉语的语法信息相对于印欧语系语言（如英语、法语）来说更加困难一些。如果我们选用基于规则方法，比如参考文献[8]中的算法，来处理汉语的术语抽取问题，那么必须要有已经标注好的熟语料。而在汉语中，要获得经过标注的语料，首先要对生语料进行切分。汉语语料的切分是印欧语系的语料中没有的问题，它是计算语言学中一个

基础性的问题。到目前为止，在这一领域已经作了大量的工作，但是结果仍然不甚理想，1%的切分错误可能会在后继的分析加工中引起10%、20%或者更高的错误：这就意味着我们难以为术语抽取系统获得理想的输入。

那么基于统计的方法又如何呢？

基于统计的方法认为：与普通词汇相比，术语拥有不同的统计特征（比如各组成成分之间较高的联系程度），依此可以鉴别出术语。大多数基于统计的方法均关注于多字术语的抽取，主要的方式是计算各组成部分之间的联系程度[17][24][25]。

在专业术语抽取领域中，汉语尽管因为具有缺乏时态、语态等特点而不适合利用基于规则的方法，但是它的如下特点却使其适合使用基于统计的方法：汉语词汇通常是单字词、双字词、三字词、四字词。根据《现代汉语频率词典》[26]，在最常用的9000个词中，各长度词汇的比例如下［27］：

一字词	二字词	三字词	四字词	五字词
26.7%	69.8%	2.7%	0.0007%	0.0002%

表2-1 词汇分布表一

另一项研究[28]中得出的统计数据如下：

二字词	三字词	n字词$(n>3)$
75%	14%	6%

表2-2 词汇分布表二

由上可见，汉语的大多数词汇长度比较短，因此我们可以集中精力考察双字词、三字词、四字词。这对于减少统计方法的计算量具有重大意义，使得在这种算法框架下汉语相对于其他语言更具优势。到目前为止，在这方面已经有了许多探索性的工作[29][30]。

基于统计学的方法首先需要针对欲处理语料建立初步的统计信息，如字频、词频等，然后根据这些原始统计信息计算出各种统计模型所需的信息，最后依据一定的标准进行候选术语项的选取。

但是统计学方法的不足之处在于它要求对大规模语料进行大量的计算，如果不设法降低算法的计算复杂度，那么在计算能力有限的情况下，必须经过

漫长的等待才能得到最终的结果，这显然是不可接受的。因此，如何根据实际情况提出具体方法以降低计算复杂度就显得尤为重要。

我们首先依据参考文献[20]中的基于统计的方法实现了一个术语抽取系统，并运行于一定规模的语料之上，得出了一些实验数据。根据对这些数据的分析，我们发现：此时可以借助规则方法来解决前面提到的计算复杂度的问题。

基于规则的方法具有简洁、概括能力强的优点，有限条规则就可以适用于海量的语言现象，利用它来对中间结果实现充分过滤，将中间结果尽可能的压缩，这对于系统的整体实现具有非常重要的意义。

具体而言，依照计算语言学理论，基于规则的方法可分为基于形式规则的方法和基于意义规则的方法：

（1） 基于形式规则的方法：与基于统计学的方法思想一致，也是从“外形”上寻找突破口，不过是以规则的形式表示出来，不像基于统计学的方法是以统计数值表现出来。这种方法既有语法规则简洁、概括能力强的优点，又避免了对字义或词义的涉及，易于实现，可操作性强；

（2） 基于意义规则的方法：试图从字或词的意义中去寻找突破口，即形式与意义之间的同构对应。这种方法需要字义或词义明确的语料作为输入，而前面已经讨论过，自动分析字或词的意义本身就是汉语处理中的基础性难题。因此，这种方法在现阶段并不具备实现的可能性。

参考文献[6]就是属于基于形式的规则方法。中以 Internet 上（中国北方农业信息网等网站）上的最新信息为语料资源，采用并列同现、模式匹配、特征词匹配等策略，在语料中抽取农业病虫害领域词汇，并利用词与词的语义相似度对词汇噪音作进一步的提出，提高了词典质量。

因此，综合汉语的特点以及具体实现时的问题，在最终实现的 DSTES 系统中，我们尝试将基于规则的方法和基于统计的方法相结合，取长补短，使整个系统达到比采用单一方法更好的性能。

基于统计学的方法由于其简洁性、直观性而受到中外学者的青睐，已经提出了多种统计模型，有学者对它们的性能做了比较[31]。但是对于基于规则的方法、特别是针对汉语的规则方法，则研究得不多，参考文献[6]作了一定的探索性工作。我们希望能在术语自动抽取领域中充分发掘基于规则方法的潜力，以提高整个术语抽取系统的性能。

2.3 系统介绍

经过前节的讨论，依据由此得出的设计思想，我们实现了一个专业领域术语自动抽取系统 DSTES。该系统的整体框架如下图所示。

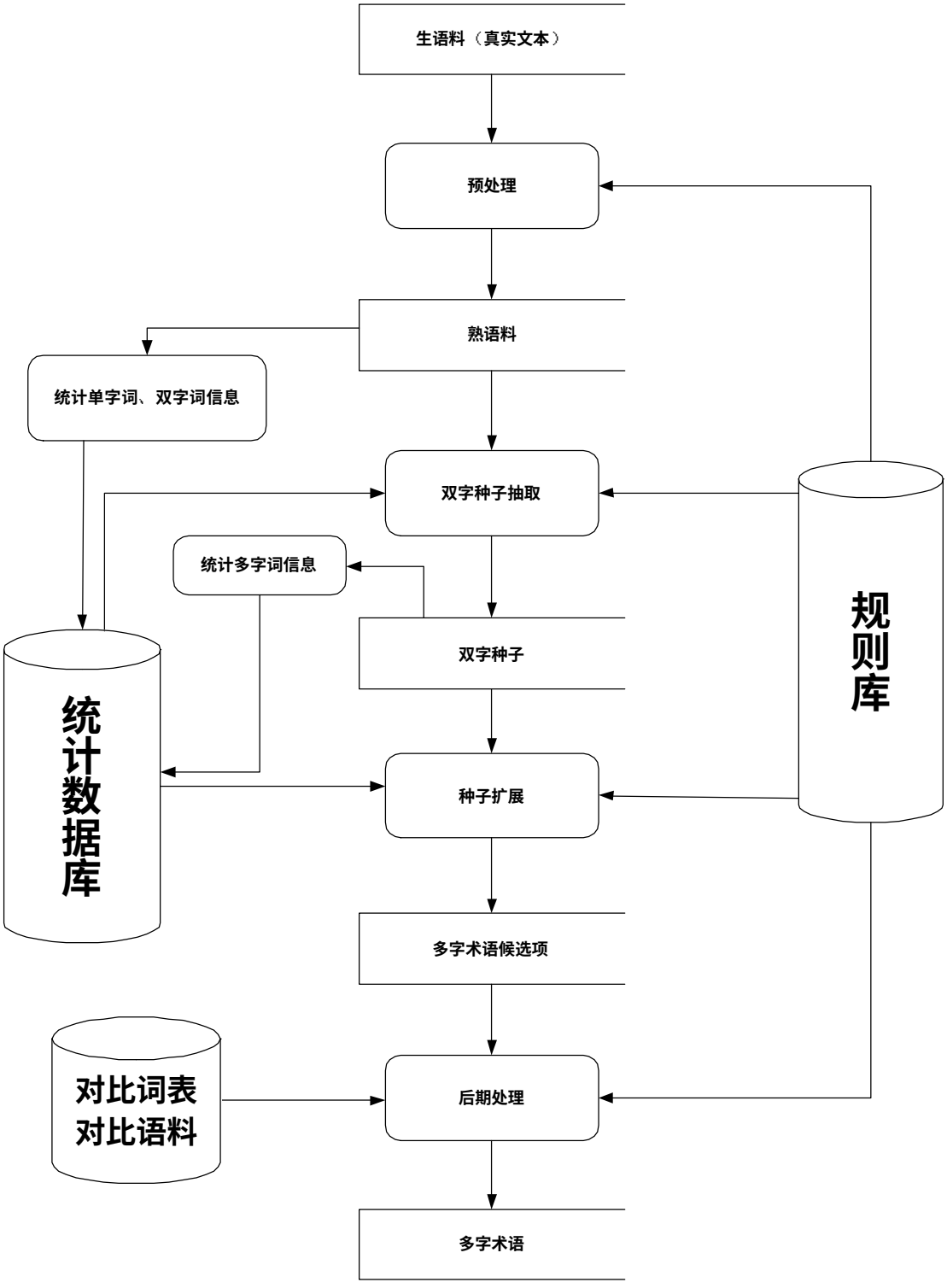

图 2.1 DSTES 系统结构图

下面分别说明 DSTES 系统各组成模块。

2.3.1 预处理模块

本模块的主要功能是对系统的输入——生语料（真实文本）进行预处理，生成符合系统特定格式和带有系统自定义标记的文本，使其方便后续模块的处理。模块具体描述如下：

输　入：	生语料（真实文本）
输　出：	符合系统后续处理要求的熟语料库 Corpus
操　作：	将独自成篇的生语料文件进行合并，获得一个语料库文件 Corpus，并在其中保留篇章信息，即在语料库 Corpus 中可以区分来自不同篇章的段落。

图 2-2 预处理模块功能说明

未经任何处理的生语料一般是独自成篇的，其存在形式是一个一个的文件。在系统的后续操作中，需要频繁的处理文件内容，如果针对单个文件分别进行存取，由于文件操作比内存操作缓慢许多，两者相差 1~2 个数量级，系统的效率可想而知，是无法接受的。如果语料存在于一个文件中，那么打开、关闭文件的操作就可以大量减少，这对于提高系统的性能是大有裨益的。所以，我们需要先对这些生语料文件进行处理，将它们合并到一个较大的文件中，我们称合并而得的文件为语料库文件 Corpus。

但是，系统在后续操作中，需要用到此处的篇章信息（见第三章投票策略）。如果这里仅仅将各段语料合并、不保留篇章信息，无法确认任意两个段落在合并前是否在同一篇生语料中，那后续的投票过滤模块将无法运行。因此，必须在语料库文件 Corpus 中保留篇章信息。我们的做法是在来自前一篇章的最后一个段落和来自后一篇章的第一个段落之间加上系统自定义的标记，该标记可以视为伪文件结束符，标志着一个篇章的结束和下一个篇章的开始。这样，来自不同篇章的段落就得到了很好的区分。

2.3.2 双字种子抽取模块

本模块的主要功能是统计单、双字词汇的词频，并初步建立统计数据库。同时，利用过滤词表对经过预处理模块处理的熟语料 Corpus 做进一步的操作。之后，从中抽取符合统计标准的双字种子，这是后续扩展操作的基础。

我们首先引入如下统计量：

定义 2-1　互信息（Mutual Information，*mi*）

对于词汇 x, y，它们的互信息 $mi(x,y)=\log_2 \frac{P(x,y)}{P(x)P(y)}$

其中，$P(x,y)=\frac{C(x,y)}{C(*,*)}$，$P(x(y))=\frac{C(x(y))}{C(*)}$），$C(x)$、$C(y)$、$C(x,y)$ 分别为词汇 x、y、xy 在语料中的词频；*为通配符，$C(*)$、$C(*,*)$ 分别为语料中单字词汇、双字词汇的个数。

定义 2-2　Log-likelihood($\log L$)[32]

对于词汇 x, y，它们的

$$\log L(x,y)=ll(\frac{k_1}{n_1},k_1,n_1)+ll(\frac{k_2}{n_2},k_2,n_2)-ll(\frac{k_1+k_2}{n_1+n_2},k_1,n_1)-ll(\frac{k_1+k_2}{n_1+n_2},k_2,n_2)$$

其中，$ll(p,k,n)=k\log(p)+(n-k)\log(1-p)$，

$k_1=C(x,y), n_1=C(x,*), k_2=C(\neg x,y), n_2=C(\neg x,*)$，$\neg$ 为取非符号，$\neg x$ 表示非 x 的词汇，其余的表达式与定义 2-1 中相同。

定义 2-3　S[20]

结合定义 2-1 与 2-2，我们得到新的统计量如下

对于词汇 x, y，有 $S(x,y)=\begin{cases}\log L(x,y) & \text{如果 mi(x, y)} \geq \text{minMutInfo} \\ 0 & \text{如果 mi(x, y)} < \text{minMutInfo}\end{cases}$

其中 $\min MutInfo$ 为阀值。

模块具体描述如下：

输　入：经过系统预处理的熟语料 Corpus，单、双字过滤词表，规则库 RDB（Rules Database）

输　出：双字种子词列表 CanList，统计数据库 DB

第一步：统计双字词频，初步建立统计数据库 DB，并记录它们在 Corpus 中的出现位置。同时，利用双字过滤词表，对 Corpus 进行过滤：对于双字过滤词表中的任意一个双字词汇 W，将 Corpus 中出现的所有 W 均以同一个特殊字符替换，表示该位置的词汇属于过滤词表，应该被过滤掉；

第二步：统计单字词频，继续扩充统计数据库 DB，并仿照第一步中的操作，利用单字过滤词表对 Corpus 进行单字词的过滤；

第三步：对双字词频进行修正，即对于 Corpus 中的所有四字词 *wxyz*，如果有下列条件成立：

$$mi(x,y) < mi(w,x) - k \text{，或者}$$

$$mi(x,y) < mi(y,z) - k$$

其中，k 为设定的阀值，则将数据库中 *xy* 的频率减 1；

第四步：对于 DB 中的双字词 xy，如果符合下列条件：

$$C(x,y) > minCount$$
$$S(x,y) > minLogL$$

其中，$\min Count$、$\min LogL$ 为设定的阀值，则将 *xy* 列入双字种子词列表 CanList 中；

第五步：对于 CanList 中的每一个双字词 xy，利用规则库 RDB 中的规则进行过滤；经过过滤的双字词保留在双字种子列表 CanList 中作为本模块的输出。

图 2-3 双字种子抽取模块功能说明

第一步和第二步利用单字过滤词表和双字过滤词表对预处理后的语料库文件 Corpus 进行了过滤。对于在这两个词表中出现的任意单字词或双字词 W，将 Corpus 中出现的所有 W 均以一个系统自定义的特殊字符替换，表示该位置的词汇属于过滤词表，已经被过滤掉。这样，在后续的操作中，只需要判断某一位置的词汇是否为该特殊字符即可，不需要再去过滤词表中进行查找。

这里将双字词的过滤安排在单字词之前进行，是基于如下考虑：一个双字词由字 A、字 B 构成，有可能存在这样一种情况：双字词 AB 在双字过滤词表中；A（或 B）在单字过滤词表中，而 B（或 A）不在。这种情况下，如果先进行单字过滤，那么会将 A（或 B）过滤掉，得到&B（A&）（&代表系统自定义的代表过滤词汇的特殊符号）。之后的双字过滤就无法将剩下的 B（或 A）过滤掉。显然，这没有实现充分过滤：剩下的 B（或 A）应该也被过滤掉。这一步骤的过滤会对后续的处理产生重大的影响，因此必须充分。显然，只需要先进行双字词过滤、再进行单字词过滤，就可以轻而易举的解决这个问题。

我们在此可以考虑对语料库 Corpus 遍历两次以完成第一步和第二步中的操作：第一遍同时进行双字词频统计和双字词汇过滤，第二遍进行单字词频统计和单字词汇过滤。这样就可以减少遍历语料的次数，从而提高系统的性能。不过在第四章中我们可以看到，与系统的主要开销相比，这里的开销相对较小。因此，此处的改进只能算是修修补补，并不能解决主要问题，无法为系统的整体性能带来质的飞跃。

第三步中对字频进行修正，其思想比较直观，就是对于四字词 $wxyz$ 而言，如果 x 与 w（或 y 与 z ）的关联程度较 x 与 y 的关联程度高（以 $mi(x,y) < mi(w,x) - k$ 来判断）（或以 $mi(x,y) < mi(y,z) - k$ 来判断），则很可能 x 与 y 分属不同的词汇：wx 属于前一个词，yz 属于后一个词。因此，对 xy 的频率减 1 以反映并强化这种推断。

要执行这个操作，必须先遍历语料库 Corpus，对四字词进行统计，然后再根据统计数据库中的信息做出调整；调整的时候还要注意不能使 $C(x,y)$ 降低到等于 0 或者小于 0，否则在后面计算统计量时会发生溢出。现在的问题是：付出这样的代价，是否值得？

我们的实验观察到的现象是：这样会对频率在 $\min Count$ 左右的候选项产生影响，使得这些词汇在修正后被排除在 CanList 之外。这些词汇大多数是非专业词汇，就它们而言，这种修正操作是完全正确的。但是这其中也有一部分是专业词汇，只是由于语料规模与分布等方面的原因，导致频率较低。这样，修正后会将这些低频术语也过滤掉。解决这个问题的一个方法就是保证语料的规模与均匀分布，使得术语的频率不至于太低。另外，根据我们的统计模型，这种修正对后续统计值的计算也会有一定的影响，使得系统正确率得到提高。

同时，我们在系统中设置了一些时间采样变量，它们记录了各个模块的开始、结束时间。通过它们可以知道，与第二步中的操作类似，这里的操作同样只是系统开销的很小一部份。

因此，综合来看，这里对字频的修正操作是值得的。

第四步中使用了一个统计量：$S(x, y)$。它基于 x 和 y 的词频信息，结合了两个统计量：互信息 $mi(x, y)$ 和 $\log L(x, y)$。对于它，我们将在第四章中专门论述，这里就不再赘述；

第五步中我们利用了多层过滤模型对双字候选项进行过滤。之所以要在这里进行过滤，是因为后续的扩展操作是以此处的双字候选项为基础的。一个双字候选项 xy 能够扩充出的若干个多字词汇，在语料量很大的情况下，这个数字是颇为可观的。关于这一点，此处只作定性说明，在第五章中将有定量讨论。如果能够在这里对双字候选项进行压缩，利用特定的过滤规则将某些不可能扩展出专业领域术语的候选项，比如金融领域语料中经常出现的“万元”，过滤掉，则系统中间结果得到了压缩，后续处理模块所需处理数据大为减少，可以大幅提高系统的整体性能。我们实验的结果也证实了这一点。

2.3.3 双字种子扩展模块

经过双字种子抽取模块的操作，我们从熟语料库中得到了双字种子列表 CanList。本模块的功能主要是针对该种子列表，对统计数据库 DB 进行一定的扩充，为后续的判断提供依据。之后对种子列表中的每一个双字种子 xy，均执行扩充操作，得到候选多字术语项。模块具体描述如下：

输　入：	双字种子列表 SeedList，统计数据库 DB，规则库 RDB
输　出：	多字术语候选项列表 TermCanList，扩充后的统计数据库 DB
第一步：	根据双字种子列表 SeedList 对统计数据库进行扩充，具体过程为：对于 SeedList 中的每一个种子词 xy，以 xy 为核心向后进行 K 个字的扩展，也就是包含 xy 的 3 字词、4 字词 … $K+2$ 字词，将扩展所得的词汇加入数据库 DB 中，计算它们的统计量，然后对 DB 中频率过低的多字词予以排除；

第二步：递归扩展双字种子词，具体过程为：对于双字种子词 xy **，考察与其相邻的每一个单字** z **（** z **可能在** xy **之前，即** zxy **；也可能在** xy **之后，即** xyz **）。如果满足下列条件：**

$$S(z,xy) > S(x,y) - k$$

其中， k **为阀值，则将** z **记录下来。在考察了所有与** xy **相邻的单字后，得到一个满足上式的列表** AdjList**。将** AdjList **中的单字根据** $S(z,xy)$ **进行降序排列，然后再顺序考察** AdjList **中的每一个单字** z **，如果** zxy **（或** xyz **）同时满足下列三个条件：**

（1） **不被** TermCanList **列表中的任何一个多字术语候选项包含；**

（2） **不能按照前述的步骤进行扩展；**

（3） **按照规则库进行的过滤无法将其排除**

则将其记入 TermCanList **表中输出。**

图 2-4 双字种子扩展模块功能说明

从双字种子抽取模块得到的双字种子列表，其中包含的双字词汇有三种：

（1） **本身就是术语，如“上市”、“基金”、“收购”等；**

（2） **只是术语的片断，如“蓝筹”（“蓝筹股”）、“盈率”（“市盈率”）、“监会”（“证监会”）等；**

（3） **既不是术语、也不可能出现在术语中，如“获得”、“随着”、“整个”等。**

尽管我们在双字种子抽取模块中已经对双字种子进行了过滤，但是由于过滤操作不可能尽善尽美，所以有了上述第三种双字词汇的存在。在后续处理中，仍然要随时考虑这些词汇的过滤，详情请见第三章中的相关讨论。

在双字种子扩展模块中，我们针对的主要是第二种双字词汇，即作为术语片断的双字种子。由于词汇可以看作其本身的片断，所以上述的第一种双字种子也可以归入第二种中一并考虑。

既然是片断，为了获取整个术语，我们就需要根据这些种子来按图索骥。因此，考察语料库 Corpus 中双字种子出现的“环境”，就成了必然的选择。

那么如何来选取这个“环境”呢？

定义 2-4　环境字符串（Environment）

设双字种子 C 在语料中的一处出现为

$$\ldots\ldots x_x x_2 \ldots x_{k-1} x_k C y_1 y_2 \ldots y_{k-1} y_k \ldots\ldots$$

在 DSTES 系统中，我们选取 C 之后的 K 个字 $y_1 y_2 \ldots y_{k-1} y_k$，加上种子本身，总共是 $K+2$ 个字，构成 C 在该处的“环境” $E(C)$，即

$$E(C) = C y_1 y_2 \ldots y_{k-1} y_k$$

对于 $E(C)$，包含种子的子串共有 $K+1$ 个。

例2-1　金融领域语料中的环境字符串

在我们处理的金融语料中有如下片断：

“…这些资金对于蓝筹股一直报有持久的青睐…”

对于双字种子“蓝筹”，如果选取 $K=4$，则该种子在此处的“环境”为字符串“蓝筹股一直报”，长度为 $4+2=6$。我们可以得到 $4+1=5$ 个子串如下：

“蓝筹”　　“蓝筹股”

“蓝筹股一”　“蓝筹股一直”

“蓝筹股一直报”

如果 K 足够大，就可以认为，我们要找的完整的术语就在这些子串中。当然，由前面的分析可知，汉语的特点决定绝大多数汉语术语并不长，所以这里的 K 并不需要太大。

所以，我们的思路就是：从这 $K+1$ 个子串之中去寻找完整的专业领域术语。

要完成这个任务，我们必须要对统计数据库 DB 进行扩充，因为数据库 DB 中到目前为止只有单字词和双字词信息，而现在我们考虑的对象要扩大到长度为 $K+2$ 的字符串，这些字符串的信息是必不可少的。所以，双字种子扩展模块的第一步操作就是扩充数据库，将长度大于 2 的字符串的信息加入进来。由前面的分析可以知道，我们只需要考察双字候选项的“环境”字符串的子串即可，对语料中所有的多字项都进行统计是不现实的（长度为 N 字的语料，其子串个数为 2^N，在 N 动辄成千上万的情况下，这显然是个天文数字），也完全没有必要。

因此，在本模块第一步操作中，我们从双字候选项进行扩充，考察且仅考察它的“环境”字符串及其子串，将所得的多字字符串加入统计数据库中，并计算它们的统计信息，具体的计算过程参见第四章的讨论。

但是，这些扩展出来的多字词，大多数出现频率过低。从统计意义上来说，在专业领域语料中，这些频率过低的多字词不可能是术语，也不可能由其扩充成术语，所以对后续操作不具备任何意义，可以将其剔除以缩小统计数据库 DB 的规模、提高查询效率。

做好统计数据库 DB 的扩充工作之后，第二步就是从双字种子的“环境”字符串的上述子串中，挑选出最有可能是完整术语的子串。这一步骤的思想是这样的：从双字种子开始，以“滚雪球”的方式将现有字符串逐步扩大，直到不能扩展为止。那么，如何判断扩展过程应该继续下去还是应该就此打住呢？也就是说：如何判断与现有多字词相邻的字是否应该被“滚入雪球中”、作为后续扩展的基础？

首先，一个先决的条件就是要满足下面的不等式：

$$S(z,T) > S(T_1,T_2) - k$$

其中T为现有待考察的多字项；T_1、T_2为T的组成部分，即$T_1 + T_2 = T$；z为与T相邻的单字（z可以在T前面，即zT，也可以在T后面，即Tz）；k为预先设定的阀值。如果满足这个不等式，则说明z与T结合得更为紧密，应该考虑将zT（或Tz）视为一体，然后在zT（Tz）的基础上继续考察。

另外，上面的模块功能说明中还给出了z是否应该与T结合在一起的三个附加条件，现在分别说明如下。

第一个条件是要判断现有的多字项W是否已经被多字术语候选列表 TermCanList 中的某个术语候选项T所包含。如果已经被包含，则不需要再扩展W了，因为即使扩展之后得到的结果也就是T。比如候选双字种子列表 SeedList 中同时包含“市盈”与“盈率”。由“市盈”扩展出专业术语“市盈率”后，再处理“盈率”时，由于它被“市盈率”所包含，即使进行扩展，所得结果也是“市盈率”，所以就不需要再进行扩展操作了。

第二个条件是不能按照前述的步骤进行扩展，也就是说，以现有待考察的多字项 T 继续进行扩展操作，如果成功，则证明 T 也为某个术语的片断，因此不能加入 TermCanList 中。显然，这是一个递归的过程。一旦牵涉到递归过程，其效率就是一个不得不考虑的因素。但是根据我们的实验结果，这里所需要的处理时间仍然不是系统的主要开销，相关讨论见第五章。

第三个条件是该多字项必须符合一定的过滤规则，这是一个核心条件。在前面的例子中我们可以看到，一个“环境”字符串可以扩充出$K+1$个待考察子

串，这些子串中绝大多数都不是专业术语或专业术语片断。对于它们，首先可以利用前面提到的频率标准进行排除——一个多字项必须具备一定的词频才有可能是术语或术语片断。但是这种操作显然是不够的。在我们的一次实验中，发现双字种子词“做空”有如下“环境”：

“…所谓做空机制，…”

根据上述的扩展方法，会有如下的扩展序列：

“做空”　　“做空机”　　“做空机制”　　“做空机制，”

到这里，我们得到了一个包含标点符号“，”的多字串。这个多字串显然不可能是术语，也不可能是某个术语的片断。尽管它的出现频率高于此处设定的过滤阀值，但是由于“，”的存在，我们在这里就可以将它排除在外，向右的扩展到此结束。

所以在第三步中，我们还需要利用另外的规则来对多字项进行筛选，只有通过筛选的多字项才能进入后续处理。具体的筛选操作见第三章的讨论，这里不再赘述。

双字种子扩展模块不仅从功能上来说是系统的核心，从占用资源上来说也是最多的。在多次系统实验中，根据我们的记录，本模块都占用相当比例的处理时间。所以，这是提升整个系统性能的关键。并且尽管在第二步中有递归操作，但是本模块的主要时间仍然消耗在第一步的扩展操作中。这主要是由于在大规模语料中，许多双字候选项都出现了成百上千次，也就有了成百上千个“环境”字符串，每个“环境”字符串又会引入$K+1$个子串，所以系统需要处理的数据量随着语料规模的增大急剧上升。关于这个问题，这里只做定性说明，详细的定量讨论见第四章。

2.3.4 后期处理模块

经过双字种子扩展模块的处理，我们得到了多字术语候选项列表 TermCanList。但是这个候选项列表中仍然包含了大量的非术语词汇，如果不予以排除，将严重影响系统整体性能。因此，我们还需要后期处理模块作最后的处理。模块具体功能描述如下：

输　入：	多字术语候选项列表 TermCanList，统计数据库 DB，语料库 Corpus，对比词表 ParaWordList，对比语料 ParaCorpus
输　出：	多字术语列表 TermList

第一步：利用统计数据库 DB，**处理术语互相包含的情况：**

对于 TermCanList **中的任意两个多字术语候选项** T_1、T_2，**记** T_1 **和** T_2 **在语料库** Corpus **中的词频分别为** $C(T_1)$、$C(T_2)$。

如果 T_1 **包含** T_2：

（1） **如果** $C(T_1)/C(T_2) < ratioThresh1$，**则将** T_1 **从** TermCanList **中排除，保留** T_2；

（2） **如果** $C(T_1)/C(T_2) > ratioThresh2$，**则将** T_2 **从** TermCanList **中排除，保留** T_1。

这里， $ratioThresh1$、$ratioThresh2$ **为预先设定的阀值，且有：**

$$\begin{matrix} ratioThresh1 \to 0 \\ ratioThresh \to 1 \end{matrix};$$

第二步：利用对比词表 ParaWordList **和对比语料** ParaCorpus，**排除通用词汇：**

对于 TermCanList **中的任意一个多字术语候选项** T，

A. **如果** T **在对比词表** ParaWordList **中，则根据词表中对** T **的分类进行判断：如果** T **属于当前处理的领域，则将** T **保留在** TermCanList **中，否则将** T **剔除；**

B. **如果** T **不在对比词表** ParaWordList **中，则统计对比语料** ParaCorpus **中** T **的出现次数** $C(T)$，**如果有**

$$C(T) > countThresh$$

这里， $countThresh$ **为预先设定的阀值则将** T **从** TermCanList **中剔除，否则将其保留。**

第三步：根据固定模式匹配，排除 TermCanList **中的日期词汇、货币词汇：**

对于 TermCanList 中的任意一个多字术语候选项T，

A. 检查T中是否包含预先建立的关键词表中的字，这些字包括“年”、“月”、“日”等表示时间的词，也包括“百”、“千”、“万”等表示数字的词，还包括“元”等表示货币的词。如果包含，则看T中这些关键词的前后是否为数字——0、1…9 或者一、二…九。如果的确是数字，则判断其为日期词汇或者货币词汇，将其排除；

B. 判断T是否为时间词汇，比如“周一”、“周二”、“周末”等。如果是，也要将其从 TermCanList 中排除。

第四步：利用语料库 Corpus，根据投票策略排除人名、地名等专有名词：

对于 TermCanList 中的任意一个多字术语候选项T，统计它共在语料库 Corpus 的多少篇章中出现过，记该篇章数为$V(T)$，如果有：

$$V(T) < \min Vote$$

这里，$\min Vote$为预先设定的阀值，则将T从 TermCanList 中剔除。

第五步：经过上述的过滤操作后，多字候选项列表 TermCanList 中剩余的多字项拷贝至多字术语列表 TermList 中，作为系统的最终处理结果输出。

图 2-5 后期处理模块功能说明

本模块的第一步中，我们针对的是 TermList 中多字术语候选项互相包含的情况，这种情况可细分为以下三种：

（1） 多字项T_1包含多字项T_2，且T_1出现的次数远比T_2少，即

$$C(T_1)/C(T_2) < ratioThresh1$$

这里，$ratioThresh1$是一个趋近于 0 的正数。这种情况表示T_1只是包含专业术语T_2的一种惯常用法罢了，其本身并不是术语。所以，应该将T_1过滤掉。在我们的一次实验中，有如下数据：

T	$C(T)$
股票	33
A股流通股股票	3

由于 $3/33 \to 0$ **,所以应该将多字项“A股流通股股票”过滤掉，保留“股票”；**

（2） **多字项** T_1 **包含多字项** T_2 **，且** T_1 **出现的次数与** T_2 **基本一致，即**

$$C(T_1)/C(T_2) > ratioThresh2$$

这里， $ratioThresh2 \to 1$ **。如果** T_1 **与** T_2 **的频率相同，则毫无疑问** T_1 **更为完整，** T_2 **应该被排除；如果** T_1 **的频率略小于** T_2 **，则表示** T_2 **的几乎每一次出现都在** T_1 **中，其余的出现可能是噪音，因此可以判断** T_1 **是专业术语，** T_2 **应该被过滤掉。**

（3） **如果上述的两种情况都不满足，则同时保留两个多字项。这种情况下，对于何者为专业术语、何者不是并无定论。由于 DSTES 系统的定位是在保证正确率的情况下提供尽可能多的候选多字术语，所以我们将两者均予以保留。我们的另外一次实验得出的数据如下：**

T	$C(T)$
国有股	72
国有股减持	22

根据前述考虑，我们将“国有股”和“国有股减持”均保留。

在本模块的第二步中，我们利用对比词表 ParaWordList **和对比语料** ParaCorpus**，排除通用词汇。所谓通用词汇，就是有些多字项，如“记者”、“中国”、“意见”等，尽管其本身是具有完整意义的词汇，但是它不属于当前处理领域，对后续的使用无甚意义，所以应该将其排除掉。**

在这里，我们使用了一个对比词表 ParaWordList**，它包含了** 31732 **条纪录，并为其中的每一个词的每一个义项都标注了所属领域。对于** TermCanList **中的每一个多字项** T **，我们先在** ParaWordList **中进行检索。如果能够检索到，证明** T **是通用词汇，需要考虑将其过滤。**

但是这个时候，必须考虑一种例外情况，即这个词语的确是当前考虑领域的专业词汇，但是比较常用，因此成为了通用词汇。金融领域中的“股票”、“股份”、“证券”，足球领域中的“后卫”、“点球”、“角球”等都是这样的例子，这类词汇显然应该被召回。

这个时候，就可以利用 ParaWordList 中的词义信息来进行判断：遍历T的每一个义项，查看它所属的领域是否为当前处理领域，如果是，则可以确定T应该被召回。

如果在对比词表 ParaWordList 中找不到T，则没有现成的词法信息可供利用。此时，可以借助一个对比语料 ParaCorpus 来进行考察。对比语料 ParaCorpus 经过筛选，不包括当前考察领域的语料，因此考察T在 ParaCorpus 中的使用情况，能从一个侧面较好的反映T的使用情况，即它是否为通用词汇。详细的讨论见第三章。具体的做法是统计T在 ParaCorpus 中的词频$C(T)$，如果有

$C(T) > countThresh$（$countThresh$为预先设定的阀值）

则说明T在非本领域的语料中也有相当程度的使用，因此T不能算是本领域的专业词汇，应该被过滤掉。

本模块的第三步利用了一些词形模式进行匹配，来判断多字项是否属于某些应该被排除的词类，如表示时间的词汇“2002 年底”、表示货币的词汇“亿元人民币”等。这些应该被排除的词类都是我们通过反复实验总结出来的，它们有些和具体的领域有关，比如金融领域中，表示货币的多字项就相对较多；有些和语料种类有关，比如在新闻语料中，有关时间的多字项就相对较多。

我们在进行词形模式匹配时，采取的策略比较简单：仅仅当该多字项符合模式“数字+关键词”时，才将其过滤掉。此处的关键词分为如下三类：

A. 时间词汇，如“年”、“月”、“日”；

B. 数量词汇，如“亿”、“万”、“千”、“百”、“十”；

C. 货币词汇，如“元”。

这是一种比较谨慎的做法，指导思想是尽可能多得保留多字候选项，对于不能肯定应该排除掉的多字项予以保留。当然，还有其他的词形模式可供利用，这有待于我们在下一步工作中继续进行归纳总结。

经过前面的过滤之后，在多字术语候选项列表 TermCanList 中，仍然存在人名、地名等专有名词，本模块第四步的任务就是处理这些多字项。这些多字项的一个共同特点是它们只在少数几篇语料中出现。而在语料规模得到保证的情况下，专业领域术语应该在语料中均匀分布，也就是说，它应该在多个篇章中出现。所以，我们可以统计多字候选项 T 出现的篇章数$V(T)$，如果有下式成立：

$V(T) < \max Count$ ($\max Count$ 为阀值)

则说明T 的出现局限在少数篇章中，它应该被过滤掉。具体操作时，尽管现在我们的语料库 Corpus 只在一个文件中，但是预处理模块已经通过增加自定义的篇章标记保留了原始语料中的篇章信息，因此我们可以利用这些篇章标记来还原原始语料并统计$V(T)$，并依照上式进行筛选。

经过前述步骤的处理，TermCanList 中剩下的多字项就是系统的最终处理结果，本模块第五步将其拷贝入词表 TermList 中输出。

2.4 系统数据结构介绍

本系统的重点在于计算，因此数据结构的设计较为简单。由于需要快速存取数据，所以我们采用数组的形式实现统计数据库 DB。

系统中用到的核心数据结构有以下三个：

（1） 词汇位置结构 CPos：用以记录单字词汇和多字词汇在语料库中的出现位置，其组成如下：

nIndex	nPos	pNext

其中，

- ✧ nIndex 为整型变量，表示该结构记录的是词汇的第几次出现；
- ✧ nPos 为整型变量，表示词汇在语料库中的开始位置；
- ✧ pNext 为 CPos 型的指针，用以指向下一个记录词汇位置的结构。

（2） 单字词 CUniFreq：用以记录单字词汇的统计信息，其组成如下：

sTerm	nCount	pPosHead

其中，

- ✧ sTerm 为字符串型变量，记录该单字词汇；
- ✧ nCount 为整型变量，记录 sTerm 在语料库中出现的总次数；
- ✧ pPosHead 为 CPos 型的指针，用以指向一个由 CPos 型结构组成的链表，该链表记录 sTerm 在语料库中出现的所有位置。

（3） 多字词 CBiStat：用以记录多字词汇的统计信息，其组成如下：

sTerm1	sTerm2	nCount1	nCount2	nCount
fMi	fLogL	fS	pPosHead	

其中，

- ✧ sTerm1 为字符串型变量，表示该多字词汇的前半部分；
- ✧ sTerm2 为字符串型变量，表示该多字词汇的后半部分；
- ✧ nCount1 为整型变量，记录 sTerm1 在语料库中出现的次数；

- **nCount2 为整型变量，记录 sTerm2 在语料库中出现的次数；**
- **nCount 为整型变量，记录整个多字词汇，即 sTerm1+sTerm2 在语料库中出现的次数；**
- **fMi：为浮点型变量，记录 sTerm1 与 sTerm2 的互信息 Mi，即** $fMi = mi(sTerm1, sTerm2)$；
- **fLogL：为浮点型变量，记录 sTerm1 与 sTerm2 的** $\log-likelihood$，**即** $fLogL = \log L(sTerm1, sTerm2)$；
- **fS：为浮点型变量,记录记录 sTerm1 与 sTerm2 的** S **值，即** $fS = S(sTerm1, sTerm2)$
- pPosHead **为** CPos **型的指针，用以指向一个由** CPos **型结构组成的链表，该链表记录** sTerm1+sTerm2 **在语料库中出现的所有位置**。

第三章 多 层 过 滤

前面已经介绍过，过滤操作对于专业领域术语的抽取具有重要意义，经过我们的多次实验，证明它能显著提高整个抽取过程的效率以及最终结果的准确性。在我们的特定领域术语抽取系统 DSTES 中，我们利用了多层过滤模型，根据多种过滤策略对系统的各个中间结果以及最终结果实现了多层过滤，取得了较好的效果。本章中，我们将详细介绍其中使用的主要过滤技术。其中，第一节介绍了利用符号和词类实现的过滤，第二节介绍了根据领域相关性理论和领域一致性理论实现的过滤，第三节介绍了根据模板匹配进行的过滤。

3.1 符号和词类过滤

在由真实语料组成的语料库中，不论何种语言，标点符号是必不可少的，占有大量的篇幅；同时会有一定数量的特殊记号，如数学运算符“∑”、“∵”，几何图形符“●”、“◆”，制表符“├”、“┘”等。我们把这些上述这些标点符号和特殊记号统称为“符号”。根据第二章中论述的双字种子抽取及其扩展算法，我们会得到相当比例的包含符号的双字种子及多字项，如下例所示：

（1） “呢？”

（2） “，但是”

其中，（1）是由于“呢”为语气词，大多数时候在疑问句句末出现，与“？”的搭配组合出现频率很高，所以作为双字种子被抽取出来；（2）是由双字种子“但是”扩展而得的多字项，“但是”作为连词出现在分句句首的频率较高，所以它与句内点号“，”搭配而成的“，但是”，其词频也相对较高。显然，这种词汇在我们的系统中应该尽早且尽可能完全的予以排除。

在参考文献[20]中，提出借助标点符号来对多字词汇进行过滤，排除包含标点符号的多字项。受此思想启发，我们在 DSTES 中也引入了符号过滤，取得了较好的效果。另外，在实验所得数据中，我们发现存在如下类型的多字项：

（1） “但股市泡沫”

（2） “随着股价”

（1）中专业术语“股市泡沫”前有连词“但”，（2）中专业术语“股价”前也有介词“随着”，这样的多字项可以肯定不是专业术语，也不可能是某个专业术语的片断。由此，我们可以从借助标点符号进行过滤的思路进行扩展，借助词类进行过滤。下面我们先讨论符号和词类方面的理论问题，在此基础上介绍我们的 DSTES 系统的具体做法。

3.1.1 符号

标点符号是辅助文字记录语言的符号，是书面语的有机组成部分，用来表示停顿、语气以及词语的性质和作用[33]。

常用的标点符号有 10 种，根据《标点符号用法》中的分类，分为点号和标点两大类，下面分别举例说明：

（1） 点号：作用在于点断，主要表示说话时的停顿和语气。点号又分为句末点号和句内点号。

A. 句末点号用在句末，表示句末的停顿，同时表示句子的语气。这种点号包括句号“。”、“.”（这种形式一般用于科技文献中）；问号“？”；叹号“！”，共计 3 种；

B. 句内点号用在句内，表示句内的各种不同性质的停顿。这种点号包括逗号“，”、顿号“、”、分号“；”、冒号“：”，共计 4 种；

（2） 标点：作用在于标明，主要标明语句的性质和作用。常用的标点有以下 9 种：

A. 引号，分为双引号“ “”、“”“和单引号“ ‘”、“’ ”，引号里面还要用引号时，外面一层用双引号，里面一层用单引号；

B. 括号，有多种形式，比如圆括号“（”、“）”，方括号“[”、“]”，六角括号“{”、“}”，方头括号“【”、“】”；

C. 破折号“——”；

D. 省略号“.....”、“..... .....”（表示整段文章或诗行的省略）；

E. 着重号“.”；

F. 连接号“—”、“--”（占两个字的长度）、“-”（占半个字的长度）、“~”（占一个字的长度）；

G. 间隔号“.”

H. 书名号，分为双书名号“《”、“》”和单书名号“〈”、“〉”，书名号里边还要用书名号时，外面一层用双书名号，里边一层用单书名号；

I. 专名号“--”，只用在古籍或某些文史著作里面，人名、地名、朝代名等专名下面，用专名号标示。

至于特殊记号，由于其广泛性及特殊性，这里就不一一列举了，3.1.3 中将介绍对它们的处理。

3.1.2 词类

汉语中词类的划分是依据句法功能还是依据意义，一直没有一个统一的标准，这主要是由汉语自身的特点造成的。这些特点包括：

1. 词无形式标记和形态变化；
2. 现代汉语中不同历史层次的成分混杂，使词的语法功能和句法规则复杂化；
3. 词的多功能现象普遍存在，使得在缺乏形态的情况下要利用语法功能划分词类更加困难；
4. 词的切分困难，使得我们难以区分某种用法是词的用法还是一个构词成分的用法；
5. 句法结构的语法关系判定困难。由于无形态，一个词在组合中到底做什么成分难以判断，也就给以语法功能为标准划分词类带来困难。

关键的原因归结为一点，还是汉语缺乏形态。[34]

这里，我们并不必纠缠于复杂的“什么是词”的语言学争论，而只是为了讨论问题的方便，根据朱德熙的词法理论对汉语词汇进行粗略的分类。

词类是根据词的语法功能分出来的类。因此同类的词必须具有共同的语法功能。汉语的词可以分为实词和虚词两大类。从功能上看，实词能够充任主语、宾语或谓语，虚词不能充任这些成分。从意义上看，实词表示事物、动作、行为、变化、性质、状态、处所、时间等等，虚词有的只起语法作用，本身没有什么具体的意义，如“的、把、被、所、呢、吗”，有的表示某种逻辑概念，如“因为、而且、和、或”等等。实词包括体词和谓词两大类。体词的主要语法功能是作主语、宾语，一般不作谓语；谓词的主要功能是作谓语，同时也能作主语和宾语。具体情况参见下表[35]：

实词	体词	1. 名　词	水	树	道德	战争		
		2. 处所词	北京	图书馆		邮局		
		3. 方位词	里	上	里头	东边		
		4. 时间词	今天	现在	从前	星期一		
		5. 区别词	男	女	金	银	新式	高级
		6. 数　词	一	二	十	百	千	万
		7. 量　词	个	只	块	条		
		8. 代　词（体词性）	我	谁	这	那	什么	

	谓词	代 词（谓词性） 这么 那么样 怎么 9. 动 词 来 写 买 研究 10. 形容词 红 大 干净 多
	虚词	11. 副 词 很 也 已经 再 12. 介 词 把 被 从 连 13. 连 词 可是 如果 即使 和 14. 助 词 的 所 得 似的 15. 语气词 啊 吗 呢 吧
		16. 拟声词 啪 哗啦 叮叮当当 叽里咕噜 17. 感叹词 哦 哎呀 嗐

表 3-1 词类表

3.1.3 过滤操作

前面已经说明过，我们的 DSTES 系统需要对双字种子以及由此扩展而得的多字项进行过滤操作，这就要求一个预先设定的过滤词表。那么，这个词表应该由什么成分来构成呢？

过滤词表的第一种成分是符号。从上面的讨论可知，标点符号的作用是用来表示停顿、语气以及词语的性质和作用。因此，作为专业术语的多字词汇，是不可能包含任何标点符号的；同样，专业术语也不可能包含特殊符号。这样，我们就可以考虑将所有的符号加入到过滤词表中。一旦在种子抽取过程和种子扩展过程中遇到这些符号，就可以将中止抽取过程和扩展过程。

但是在实际操作时，我们必须有所取舍。这里的过滤操作用于双字种子词的抽取及种子词扩展的过程，其中扩展过程是个递归的过程，所以其效率非常重要。如果过滤词表过于巨大，那么每次操作都必须花费大量的时间，这显然是不允许的，会成为整个系统的瓶颈。但是如果词表过小，就会导致很多本应该过滤掉的多字词被保留下来，过滤操作的效果就不明显了。尽管过滤操作的效率提高了，但是后续操作需要额外处理很多本应该过滤掉的非术语多字词，这同样会降低整个系统的性能。所以，我们必须将过滤词表的规模控制在一定的范围内。

对于标点符号而言，它们的数量有限，但却是最常用的，不管系统具体处理的是什么领域，它们在相关语料库中都占有相当大的比例。所以，我们必须将它们包含进过滤词表中。另外，从 3.1.1 中可以看出，标点符号的个数只有区区数十个，所以完全可以全部包含进去而不影响系统的整体性能。

但是对于特殊符号，情况就不同了。不同领域使用的特殊符号千差万别，很难一概而论，数学领域使用的特殊符号“∫”、“≦”、“∨”与化学领域使用的特殊符号显然有天壤之别。因此，要穷尽各个领域，将各领域的特殊符号都填充至过滤词表中，是不现实的，系统效率也会大受影响。同时，这样也是没有必要的——我们只需要针对当前处理的领域，搜集该领域中使用的特殊符号即可。如果更换了处理领域，我们就可以通过更换过滤词表的这一部分来适应这种变化。当然，这样做的一个损失是失去了系统的领域无关性，系统必须根据当前处理领域做出调整了。不过，综合来看，这种损失是值得的。

过滤词表的第二种成分是词汇。我们根据表三中对汉语词汇的分类，对各类词汇分别予以考察，考察的目的是判断该类词汇是否能出现在专业领域术语当中（或者其本身就是术语）。

首先，汉语词汇分为实词和虚词两大类。实词是开放类，虚词是封闭类。所谓开放类，指的是难于在语法书里一一列举其成员的大类。所谓封闭类，是指可以穷尽地列举其成员的不很大的类[35]。参考文献[36]中收录了副词、介词、连词、助词、语气词等虚词 790 条。

因此，从实现的角度出发，我们先考察了各类虚词。很容易就可以看出，语气词（如“啊”、“吗”、“呢”）、拟声词（如“啪”、“哗啦”、“叮叮当当”）、感叹词（如“哦”、“哎呀”、“嗐”）都不可能出现在任何领域的专业术语中，因此可以将这三类词汇全部放入过滤词表中。

剩下的四类虚词——副词（如“很”、“也”、“已经”）、介词（如“把”、“被”、“从”）、连词（如“可是”、“如果”、“即使”）、助词（如“的”、“所”、“得”），这几种词汇必须谨慎处理，因为它有可能包含在某些领域的术语中。比如助词“的”，它在绝大多数情况下不会出现在专业术语中，但是在中药领域，就有“阿的松”（一种中药名）。

对于各类实词也是如此，在将其列入过滤词表时，必须慎重考虑，因为一旦将其列入，那么所有包含该词的多字项都将被过滤掉，如果有某些术语包含该词，那它们将被排除在最终结果之外。

在我们的系统中，根据词的长度对过滤词表进行了分类，分为单字过滤词表和双字过滤词表：

（1） 单字过滤词表：包含长度为 1 的过滤词汇，主要成分为

①标点符号，如“，”、“。”、“《”。我们的 DSTES 系统主要的处理对象是中文生语料库，但是有可能由于输入错误，在其中引入了“,”、“,”等英文标点符号，所以我们也将英文标点符号引入过滤词表中，使得系统能够处理这些情况；

②语气词、拟声词、感叹词，如“吧”、“嗦”。这三种词还有一部分是双字词，在下面的双字过滤词表中引入；

③其他长度为 1 的过滤词汇，如“她”、“们”、“竟”。这一部分是不在任何领域的任何术语之中出现的单字，当然，这一点是比较难以确认的，我们只能凭借有限的经验来判断，所以我们放入表中的这部分词汇相对较少。

（2） 双字过滤词表：包含长度为 2 的过滤词汇，这些词汇同样不能出现在任何领域的任何术语之中。相对于长度为 1 的词汇，这部分的确认要容易得多，所以这部分词汇要多一些。当前，我们的单字过滤表中有 198 个词汇，而双字过滤表中有 507 个词汇。当然，这些词汇只是我们在实验数据的基础上总结出来的，是远远不够的，还需要极大的扩充。

我们只在系统中设置了单字过滤词表和双字过滤词表，并没有考虑三字、四字词的过滤，这主要是因为在汉语中，单字词和双字词已经占到了所有词汇的绝大多数，根据表 2-1，在最常用的 9000 词中，单字词和双字词占的比例是 96.5%。所以，我们只需要考虑单字词和双字词即可覆盖绝大多数情况，考虑三字词不仅需要额外的处理时间，而且意义不大。

在具体实现时，我们并没有按照比较直观的做法，即在双字种子抽取和种子扩展的时候再去过滤词表中查找相应字词；相反，我们早在双字种子抽取模块的第一步和第二步中进行单字词频和双字词频统计的同时，就进行了过滤：对于遍历语料库 Corpus 的过程中遇到的每一个词，去过滤词表中进行查找。如果能找到，则将该位置的该词汇替换为同等长度的系统自定义符号，表示此位置的词汇对整个术语抽取操作没有实质意义，已经被过滤掉了。这样，我们只用了一次遍历，即解决了过滤问题。此后的操作，当需要判断某个位置的词汇是否需要过滤时，只需要判断该位置上现有的是否为系统自定义的过滤符号即可，不需要再对过滤词表进行检索。如果不做这样的改进，由于双字种子的扩展是一个递归过程，所以会对同一位置的词汇重复判断，这样显然是毫无意义的，会极大降低系统的效率。

3.2 领域相关性和领域一致性过滤

我们的 DSTES 系统的输入是汉语生语料库，输出的是特定领域的专业术语。那么，一个根本性的问题就是：何为专业术语？

参考文献[37]中指出，术语具有如下基本特征：

（1） 专业性：术语是表达各个专业的特殊概念的，所以通行范围有限，使用的人较少；

（2） 科学性：术语的语义范围准确，它不仅标记一个概念，而且使其精确，与相似的概念相区别；

（3） 单义性：术语与一般词汇的最大不同点在于它的单义性，即在某一特定专业范围内是单义的。有少数术语属于两个或更多专业，如汉语中“运动”这个术语，分属于政治、哲学、物理和体育 4 个领域；

（4） 系统性：在一门科学或技术中，每个术语的地位只有在这一专业的整体概念体系中才能加以规定。

但是，这些基本特征并不具备较好的操作性，因此我们必须转换思路。显然，“专业术语”是相对于“一般词汇”而言的，如果没有“一般词汇”的界定，那么“专业词汇”的界定也就无从谈起；反之亦然。因此，上面的问题可以归结为：专业术语与一般词汇的区别在哪里？

既然是特定领域的专业术语，那么很显然，有这样两个要求：

（1） 在当前处理领域内通用，即在当前领域的语料中广泛出现，而不是局限在某一个局部；

（2） 在其他领域中不通用，即在除当前领域之外的所有其他领域语料中很少出现或基本不出现。当然，有两种特殊情况需要考虑：

（1） 本身是专业术语，但是后来由于在各种媒介中使用频繁，逐渐被大众所接受，进入了通用领域，成为通用词汇，如“股票”、“非典”、“克隆”等；

2） 本身是专业术语，但是在某些其他领域中（一般是相近领域）也是专业术语，即该词汇是多义词，在不同领域中有不同含义。一个很好的例子就是“中锋”，它即在足球领域中使用，也在篮球领域中使用。

参考文献[37]中根据术语的使用范围，对术语做了进一步的分类：

（1） 纯术语。它的专业性最强，如“等离子体”；

（2） 一般术语。它的专业性次之，如“压强”；

（3） 准术语。它已经渗透到人们生活中，其专业性最弱，如“塑料”。

上面（2）的第 1 种特殊情况就是这里所说的准术语。

在系统实验中，我们发现有如下的一些词汇：

（1） “唐万里”（人名）、“万杰高科”（股票名）、“海尔”（公司名）。这种词汇尽管可能在语料库的某个局部出现频率很高，但是它们基本局限在语料库的某一部分，在本领域内不通用，不符合上面的条件（1）；

（2） “记者”、“国务院”、“北京”。这种词汇不仅在本领域内通用，在其他领域内也是广泛使用的。所以，它们不符合上面的条件（2），即只在本领域内通用。

因此，在系统中必须针对这些词汇进行相应的过滤操作。但是仅有过滤是不够的，我们还要考虑进行一定的召回操作。正如上面的条件（2）中提到的，还存在特殊情况，如“守门员”、“股票”、“投资”等。尽管不局限在特定领域，它们仍然是该领域的专业词汇，应该包含在最终的抽取结果中，所以要把它们召回。

在本节中，我们先概要介绍与上面条件相关的两个概念：领域相关性和领域一致性，然后介绍我们的 DSTES 系统中利用它们进行的过滤操作。

3.2.1 领域相关性和领域一致性

参考文献[38]中提出了领域相关性（Domain Relevance）和领域一致性（Domain Consensus）的概念，这是两个基于熵（Entropy）的统计量，下面我们分别予以介绍。

在语料库中，术语与非术语（如“北京”、“科技”、“报告”等）的词频都有可能很高，因此单纯依靠词频无法衡量一个词语与某个特定领域的相关程度。那么如何衡量这种领域相关程度呢？参考文献[38]认为，可以通过与不同领域进行比较分析来衡量一个术语候选项和特定领域的相关性，即“领域相关性”（Domain Relevance， DR），DR 的具体数量定义如下：

$$DR_{t,k} = \frac{P(t \mid D_k)}{\sum_{j=1}^{n} P(t \mid D_j)}$$

这里，$DR_{t,k}$ 表示术语 t 与领域 D_k 的相关性；$D_1, D_2, \ldots D_{n-1}, D_n$ 为给定的 n 个领域；$P(t \mid D_k)$ 由下式进行估算

$$E(P(t \mid D_k)) = \frac{f_{t,k}}{\sum_{t' \in D_k} f_{t',k}}$$

其中，$f_{t,k}$ 是术语 t 在领域 D_k 中的频率。

如前所述，术语必须在特定领域内的语料中广泛出现，不能局限在语料库的某个局部，也就是说本领域中对该术语的使用必须具有某种程度的一致性。由

此引入了“领域一致性”（Domain Consensus，DC）的概念，DC 的具体数量定义如下：

$$DC_{t,k} = \sum_{d \in D_k} \left(P_t(d) \log \frac{1}{P_t(d)} \right)$$

这里，$DC_{t,k}$ 表示术语 t 在领域 D_k 中的“一致性”；$P_t(d)$ 表示文档 d 包含术语 t 的概率。

参考文献[38]将 DR 与 DC 进行线形组合以衡量术语 t 是否应该被过滤：

$$DW_{t,k} = aDR_{t,k} + (1-a)DC_{t,k}^{norm}$$

其中，$DC_{t,k}^{norm}$ 为 $DC_{t,k}$ 经过规格化后的结果；$a \in (0,1)$。

从上面对 DR 和 DC 的定义中可以看出，这些统计量实际操作起来是比较困难的。因此，在 DSTES 系统中并没有按照它们的上述定义来实现对术语的过滤：我们只是借鉴了它们的名称以及其表达的思想，即衡量术语与特定领域的相关性和一致性。在系统中，通过反复实验、归纳总结，我们采用了另外一些简单易行的方法。下面就是 DSTES 系统中具体做法的介绍。

3.2.2 过滤操作

我们先介绍对领域相关性问题的解决。

根据前面的论述，领域相关性指的是术语与特定领域的相关程度；换句话说，特定领域的术语在其他领域的语料（即对比语料）中应该很少出现或基本不出现。这样，如果能够确定某个术语候选项在其他领域语料中也频繁出现，证明它与其他领域的相关程度也很高，就可以考虑将其过滤。但是，前面提到过，这样会将“股票”、“非典”等已经变成通用词汇的专业术语以及“守门员”、“中锋”等在多个领域中均为专业术语的词汇排除掉，因此还要将这些专业术语召回。

在 DSTES 系统中，我们使用了两种对比资源来确定候选术语在其他领域的使用情况：对比词表和对比语料。

我们在 DSTES 系统中采用的对比词表比较简单，其结构如下：

Word	Semantic

表 3-2　对比词表结构

其中，Word 存储词汇本身；Semantic 存储该词汇对应的语义，如果有多个义项，则用自定义的特殊标记将这些义项隔开。这个词表中包含了 31732 个词汇，

绝大多数是名词。它们来自各个领域，都是常用词。我们可以将该词表视为一个熟语料库，区别于由真实文本构成的生语料库。对于每一个待筛选的术语候选项，我们都在该词表中进行查找。如果能找到，证明该候选项是常用词，应该被过滤掉。上文提到的“记者”、“国务院”、“北京”等词，均可以通过这种方式予以排除。

当然，如上所述，利用对比词表实现的过滤必须考虑某些专业术语的召回问题，这需要考察词汇的义项。因此，我们在决定将候选术语项 t 排除之前，要对该词汇的词义进行判断，如果该词汇有属于当前领域的义项，仍然需要将该候选项予以保留。例如对比词表中，“股票”一词的词义是“coupon|票证,#fund|资金”，如果当前处理领域是金融领域，我们根据词表中对词汇的分类，确定“商”、“货币”、“价格”、“酬金”、“钱财”、“票证”、“费用”等七类词汇和金融领域相关，那么这两种义项都可以将“股票”一词召回，保留在候选项列表中。

但是这个词表中收录的词汇显然是不全的：名词是开放词类，无法一一列举其成员，三万多条只能说是沧海一粟；另外，词表中绝大多数是名词，就意味着此时只能筛选掉属于名词的常用词。还有大量的动词、形容词等，就无法在这里予以排除了。

针对这一问题，我们使用了另一种对比资源：对比语料，即未经处理的生语料。对于经过对比词表过滤的候选术语项 t，统计它在对比语料中出现的频率 $C(t)$。如果有下式成立：

$$C(t) > \max Count$$

其中，$\max Count$ 为预先设定的阀值，则表明术语候选项 t 在对比语料中也是通用词汇，不符合领域相关性的要求，所以要将其过滤掉。

对比语料选取的是和当前处理领域不相关的其他领域语料。之所以要不相关，是因为在相关领域中可能存在一些共同的专业术语（尽管这些术语在不同领域中的含义往往不同），如前文所说的“中锋”就同时存在于足球领域和篮球领域。如果选用相近领域的对比语料，则本领域的专业词汇也会在对比语料中频繁出现，最终导致该词汇被错误的过滤掉。

从上文可以看出，对比词表和对比语料都可以衡量术语候选项的领域相关性：它们的作用都是判断术语候选项在非专业领域中是否常用。不同的是对比词表是从真实语料中归纳而得的，其中全部都是常用词；而对比语料是真实文本，有常用词，也有非常用词，必须进行词频统计以判断词汇是否常用：前者可以看作后者的高级形式。它们之间的比较如下表所示：

	优点	缺点
对比词表	1、规模小，搜索代价低； 2、有语义信息，可判断词汇所属领域以进行召回操作；	1、必须由人工来构建，代价大，效率低；
对比语料	1、其实质为真实文本，只需简单筛选即可；	1、规模大，搜索代价高； 2、利用语义信息的代价较高、难度较大；

表 3-3 对比词表与对比语料的比较

接下来，我们要解决的是领域一致性问题。

如前所述，领域一致性考察的是专业术语在本领域内的使用情况，反映到语料中，就要求术语在本领域语料中均匀分布，不能局限于某个局部。为了衡量专业术语 t 在语料库中的分布情况、并利用所得结论对 t 进行过滤，我们设计了一个如下所示的“投票”策略：

$Vote(t) = 0$;

For 语料库中的每一篇章 *Do*

Begin

If 该篇章中包含 t

Then Vote(t) = Vote + 1

End

IF Vote(t) < $\min Vote$

Then 过滤 t；

Else 保留 t。

其中，$\min Vote$ 为预先设定的阀值。

表 3-4 投票策略

这种策略和社会生活中的投票选举过程有些类似，所以我们把这种策略称为“投票”策略：对于每一个候选术语项 t 来说，语料库中的每一篇章可以看作是一个选举人，握有一票。如果该篇章中出现了 t，则该篇章将手中的票投给 t，

否则不投。遍历完所有的篇章后，统计 t 所得选票数 $Vote(t)$，如果选票数大于一个预先设定的阀值 $\min Vote$，则说明 t 在语料库中的分布比较均匀，即 t 在领域中的一致性较强；反之则说明 t 局限在语料库的某个局部，它在领域中的一致性较差，应该将其过滤掉。

通过这种策略，我们成功的过滤掉了相当数量的人名、地名、机构名等不属于专业术语的词汇，如“万杰高科”、“财经时报”、“营业部”等。但是在语料规模较小、某些专业词汇只在少数篇章中出现的情况下，这样也会过滤掉“小盘股”、“商业银行”等专业术语。如果语料达到一定的规模，使得绝大多数术语都能实现均匀分布，那么这个问题就迎刃而解了。

3.3 模板匹配过滤

前节中介绍了根据术语候选项在本领域及对比领域中的使用情况来实现的过滤。经过该方法的操作，我们在实验数据中发现，仍然存在一些种类的非术语候选项无法过滤，分别举例如下：

（1）“2002 年底”、“3 月初”、“截至 3 月”；

（2）“元每股”、“亿元人民币”；

（3）“在 2003 年”、“提案中”、“会议上”。

其中，（1）是由于语料从网站新闻直接而来，并且时间相对比较集中，所以这些表示时间的词汇在语料库中词频较高，并且这些词汇在对比词表不会出现（容量有限的对比词表不可能包含这些表示具体时间的词汇），在对比语料中词频也不会太高（除非对比语料库中的语料取自同一时段）；（2）是由于实验所处理的语料来自金融领域，这些词汇在金融领域中显然是比较常见的，而它们在对比词表和对比语料中显然不会太通用；（3）是由于对比语料没有准确地反映这些词汇在对比领域中的使用情况——对比语料不可能完全覆盖所有的通用词汇。当然，“在 2003 年”也存在（1）中提到的原因，即不能苛求对比资源中准确反映这些具体时间词汇的使用情况。

既然无法利用上节中使用对比资源、考察候选术语项外部环境的方法来过滤上述词汇，那么这一节中，我们把注意力转向术语候选项本身，讨论如何利用词汇本身的信息进行过滤。

参考文献[8]中使用经过标注的文本，从中抽取了符合如下条件的多字项：

（1） 具有形式 $((A|N)^+|((A|N)^*(NP)*)(A|N)^*)N$，其中 A 为形容词（Adjective），N 为名词（Noun），NP 为名词短语；

（2） 出现频率超过一定阀值。

参考文献[6]中提出了用模式匹配的方法来进行专业词汇的获取。这些模式只是一些简单的专业词汇及其相关知识的触发抽取模式，如：

A：[防治]<>

B：<>[主要危害]

C：TriggerB[兼治]<>等 TriggerE[病害]

A、B、C 分别代表前触发型、后触发型、前后触发型，<>的内容即为所要抽取的词汇。

尽管这两种方式有所不同，但是它们的一个共同点就是利用了模板匹配的思想，不同的只是前者利用的是词汇的语法信息进行匹配，而后者利用的是简单的字符串匹配。

受此思想启发，我们根据语法理论知识，在 DSTES 系统中设计了一些模板匹配模块。不同的是，上面两种算法利用模版匹配进行术语的抽取，而我们的系统中则利用模版匹配来进行术语候选项的过滤。

我们在 DSTES 系统中使用了三种类型的模板：

（1） 时间词模板

由于在实验中我们采用的主要是来自相关网站（新浪、搜狐等）的新闻语料，其中表示时间的词汇数量较多，词频也比较高，因此必须着力研究这类非术语词汇的过滤办法。我们这里单独设计了一个模版，用以处理时间词汇的过滤：

$$Word^{*}Num^{+}TimeWord^{*}$$

模板一 时间词模板

$Num=$｛“0”，“1”，“2”，“3”，“4”，“5”，“6”，“7”，“8”，“9”，“零”，“一”，“二”，“三”，“四”，“五”，“六”，“七”，“八”，“九”，“十”，“两”，“几”，“多少”，“百”，“千”，“万”，“亿”，“来”，“多”，“好几”，“好些”，“若干”，“半”｝，它表示数量；

$Time=$｛“年”，“月”，“日”｝，它表示时间；

Word 为任意单字词汇；

这里兼顾了数字的汉字表示法与阿拉伯表示法，是因为这两种形式都是大量存在的，“1998 年”和“一九九八年”均很常见。如果候选术语项 t 符合时间词模板，则可以将其排除；

（2） 量词模板

量词是能够放在数词后头的粘着词[朱德熙，语法讲义]。通过实验，我们发现包含“数词+量词”形式子串的非术语词汇在术语候选项列表中也占有相当的比例，因此，我们设计了量词模板来处理这类词汇：

$$Word^{*}Num^{+}QuaWord^{*}$$

模板二 量词模板

其中， Num **和** $Word$ **的定义与模板一中相同；** Qua **为量词集合，由于成员数量众多，这里就不一一列举了。如果候选术语项** t **符合该模板，则将其过滤掉。但是，这里有一个原则，就是对于量词** q **，为谨慎起见，只有确信在当前处理领域中不可能出现** $Word^{*}Num^{+}qWord^{*}$ **形式的专业术语时，才可以将** q **加入** Qua **集合中。显然，这个判断是和具体领域相关的。**

（3） **介词模板**

最后，我们介绍利用介词产生的模板。从语义上看，介词的作用在于引出与动作相关的对象（施事、受事、与事、工具）以及处所、时间等[35]**。显然，包含介词的候选术语项可以确定不是专业术语。由此，我们可以总结出以下模板以排除这类词汇：**

$$\mathrm{Pr}eWord^{*}$$

模板三 介词模板

其中， $Word$ **的定义与模板一中相同；** $\mathrm{Pr}e$ **为介词集合，包含“在”、“把”、“给”之类的介词。与模板二中一样，这里也必须谨慎行事，只有确认在当前处理领域中不可能出现** $pWord^{*}$ **形式的专业术语时，才可以将** p **加入** $\mathrm{Pr}e$ **集合中。比如在金融领域中，介词“跟”不可以放入** $\mathrm{Pr}e$ **集合中，因为存在“跟进”等专业词汇。**

很显然，在这里的介词模板中，完全可以不用局限于介词：我们可以在 $\mathrm{Pr}e$ **集合中加入任何词汇** p **，只要当前领域中不存在** $pWord^{*}$ **形式的专业术语即可。例如，在金融领域中，连词“那”就可以加入** $\mathrm{Pr}e$ **集合中。**

另外，与这种前置形式的模板相对应，我们可以使用另外一种形式的模板：后置形式的模板，即：

$$Word^{*}Post$$

模板四 后置模板

与前置形式的模板类似，可在集合 $Post$ **中加入任何种类的词汇** p **，只要当前领域中不存在** $pWord^{*}$ **形式的专业术语即可。例如，可将助词“着”加**

入到 *Post* 集合中，因为在几乎所有领域中，都没有以“着”结尾的专业术语。

第四章 统 计 方 法

上一章详细介绍了特定领域专业术语抽取系统（DSTES）中使用的过滤技术及其理论背景，它们大多基于规则方法。如前所述，由于汉语自身的特点，如缺乏时态、语态词形变化等信息，使得分析并使用汉语的语法信息相对于印欧语系语言（如英语、法语）来说更加困难一些。所以，在专业术语抽取领域中，单纯依靠基于规则的方法显得比较困难。因此，我们在 DSTES 系统中引入了基于统计的方法。这一章我们就来介绍该系统中使用的统计方法。其中，第一节讨论在专业术语抽取领域中经常使用的几种统计模型，并介绍 DSTES 系统中使用的模型；第二节讨论系统中占用资源最多的步骤——双字种子的扩展操作；第三节介绍两个细节问题。

4.1 统计模型的比较

在专业术语抽取领域中提出的统计方法大多基于这样的基本思想：通过特定的统计模型来衡量候选字符串内部各单位之间的结合紧密程度，然后利用阀值型分类器进行决策，即如果得出的统计量高于一定的阀值，则将其抽取出来，经过一定的操作（如扩展、过滤）之后，作为专业术语输出。由此可见，这里使用的统计模型至关重要，它是整个系统的基石，很大程度上决定了系统整体性能的高低。那么，这些统计模型到底孰优孰劣呢？

参考文献[31]中分别考察了九种常用的统计模型在汉语专业术语抽取方面的表现，并尝试将它们组合在一起（采用遗传算法来自动调整组合权重），以提高性能。由于二字词在汉语词汇中占有举足轻重的地位（见第二章中词汇分布表），文中针对二字词进行抽取实验，具体实验数据参见表 4-1。

这里，x, y 分别表示组成二字串的单字，$\bar{x}$ 表示非 x 的字；N 表示语料库的规模；f_x 和 P_x 分别表示字 x 出现的频次和概率，f_{xy} 和 P_{xy} 分别表示串 xy 出现的频次和概率；x_{xy} 则表示在字 x 、字 y 独立的条件下 f_{xy} 的期望值，显然有：

$$x_{xy} = P_{xy}N = P_x P_y N = f_x f_y \mid N$$

由此文中得出结论：这九种模型中，互信息（Mutual Information，MI）的抽取能力最强，并且各种统计量之间并不具备良好的互补性。

尽管互信息（*MI*）较好的衡量了词汇各组成部分的结合紧密程度，但是它有一个显而易见的弱点：没有考虑词频因素。因此即使词汇 *xy* 在语料库中出现的频率（即 P_{xy}）很低，只要 *x* 和 *y* 的频率（即 $P(x)$、$P(y)$）同样低，那么 MI_{xy} 仍然会很高。所以，只依靠互信息无法处理低频情况、排除噪音。

方法	**记为**	**公式**	F-Measure(%)
Frequency	Freq	f_{xy}	26.28
Mutual Information	MI	**见定义** 2-1	54.77
Selectional Association	SA	$\frac{P(x\mid y)MI(xy)}{\sum_z P(z\mid y)MI(zy)}$, **其中** $P(x\mid y)=\frac{f_{xy}}{f_y}$	42.98
Symmetric Conditional Probability	SCP	$\frac{f_{xy}^2}{f_x f_y}$	51.77
Dice Formula	Dice	$\frac{2f_{xy}}{f_x+f_y}$	49.37
Log-likelihood	LogL	**见定义** 2-2	43.13
Chi-squared	Chi	$\frac{N(f_{xy}f_{\bar{x}\bar{y}}-f_{x\bar{y}}f_{\bar{x}y})^2}{(f_{xy}+f_{x\bar{y}})(f_{xy}+f_{\bar{x}y})(f_{\bar{x}\bar{y}}+f_{x\bar{y}})(f_{\bar{x}\bar{y}}+f_{\bar{x}y})}$	52.97
Z-Score	ZS	$\frac{f_{xy}-x_{xy}}{\sqrt{x_{xy}(1-x_{xy}/N)}}$	53.20
Student’ s t-Score	TS	$\frac{f_{xy}-x_{xy}}{\sqrt{f_{xy}(1-f_{xy}/N)}}$	39.12

表 4-1 术语抽取领域常用统计量及其性能比较

针对这一情况，参考文献[20]中综合了在低频情况下表现相对较好的另一种统计量 Log-likelihood($\log L$)，提出一种新的统计模型 $S(x,y)$（见定义 2-3）。需要说明的一点是，尽管 $S(x,y)$ 是两种统计量的组合，但是它并不是线性组合，不

在参考文献[31]讨论的范围内，所以不受其结论的影响。我们在 DSTES 系统中借鉴的就是这种统计模型。

但是，这里有一个关键问题：上述统计量都是针对二字串提出的，计算公式中的 x 和 y 定义明确，分别为二字串的前一个字和后一个字。当需要计算的字串为 $x_1x_2...x_{(n-1)}x_n(n>2)$ 时，如何选取 x 和 y？显然，$MI(x_1,x_2...x_{(n-1)}x_n)$ 与 $MI(x_1x_2...x_{(n-1)},x_n)$、$LogL(x_1,x_2...x_{(n-1)}x_n)$ 与 $LogL(x_1x_2...x_{(n-1)},x_n)$ 在绝大多数情况下是不同的，所以对 $x_1x_2...x_{(n-1)}x_n$ 作不同切分，会得到不同的 x、y 以及 $MI(x,y)$、$LogL(x,y)$、$S(x,y)$。由于字符串长度为 n，所以会有 $n-1$ 种切分。那么，哪一种切分才是正确的呢？

我们认为，上述几种统计量的最终目的都是要衡量整个字符串内部各部分的结合紧密程度，以此作为阀值型分类器的判断依据。为了不至于造成遗漏，我们应该以统计量可能达到的最大值来衡量这个字符串。这就要求我们选择一种使得统计量达到最大值的切分方式。由于本系统使用的主要统计量是 $S(x,y)$，所以我们的选择标准就是 $S(x,y)$ 的大小，即这 $n-1$ 种切分中使得 $S(x,y)$ 最大的一种切分：

$$\arg\max_{x,y} S(x,y)$$

4.2 数据库扩充

前节中介绍了 DSTES 系统中使用的统计模型，现在我们来讨论系统实现中的一些具体问题。

在本系统的实现过程中，最大的一个难点就是系统整体效率偏低。为了系统的最终实用，必须着力提高系统效率。我们在系统中设置了时间采样变量，发现系统的主要处理时间消耗在数据库的扩充过程中。在一次测试中，我们得到的各模块运行时间占系统整体运行时间的比例如下：

模块名称	运行时间
单字词频统计	3.09%
双字词频统计	8.97%
种子词汇抽取	27.31%
数据库扩充	32.21%
种子词扩展	7.69%
后期处理	20.73%

表 4-2　各模块运行时间比例

因此，如果能提高占用时间最多的数据库扩充过程的效率，那么系统整体性能就有望得到大幅改善。我们在这一节中详细讨论该过程的操作步骤，尝试通过各种方式来改进该过程。

我们所作的一个主要尝试是对双字种子的“环境”字符串的控制。在前面已经介绍过，我们这里所说的“环境”，指的是语料库中双字种子词和它之后的 K 个字组成的字符串。那么现在就存在两个问题：

1、　为什么只包含种子词后面的 K 个字，而不考虑种子词前面的字词？

2、　K 应该取什么值？

我们先来看第一个问题。在参考文献[20]中，考察双字种子的“环境”时，采用的是前后兼顾的方法，即截取种子词前后各 K 个字组成字符串作为该种子词的“环境”。显然，这里的“环境”字符串的长度为 $2K+2$，由它衍生出来的包含种子词的子串共有 $(K+1)^2$ 个。以前文的例子而言，金融领域的语料中有如下片断：

“…这些资金对于蓝筹股一直报有持久的青睐…”

对于双字种子“蓝筹”，如果选取 $K=3$，则该种子在此处的“环境”为字符串“金对于蓝筹股一直”，长度为 $2*3+2=8$。我们可以得到 $(3+1)^2=16$ 个子串如下：

“蓝筹”　“蓝筹股”　“蓝筹股一”
“蓝筹股一直”　“于蓝筹”　“于蓝筹股”
“于蓝筹股一”　“于蓝筹股一直”　“对于蓝筹”
“对于蓝筹股”　“对于蓝筹股一”　“对于蓝筹股一直”
“金对于蓝筹”　“金对于蓝筹股”　“金对于蓝筹股一”
“金对于蓝筹股一直”

这种处理方法比较直观：既然前后是对称的，那么种子词汇完整的“环境”应该包括前后文，所以应该前后兼顾，一起考察。但是，我们通过实验发现，这个扩展过程会产生大量的冗余。考察语料库中的某个片断，

$$“...C_1\underbrace{.....}_{n}C_2...”$$

其中有两个双字种子词C_1、C_2，它们之间有n个字，每个种子词可以向左右分别扩充K个字。由于C_1、C_2的长度均为2，如果$K \geq n+2$，则从C_1开始的扩展过程和从C_2开始的扩展过程会得到$(K-n-1)^2$个重复的多字项，这些多字项均包括$C_1\underbrace{.....}_{n}C_2$。如果$K$足够大，这显然是个不小的开销。

如果采取只从双字种子词开始向后扩充的策略，根据相同的推导过程可知，在$K \geq n+2$的情况下，会有$(K-n-1)$个重复的多字项。当K足够大时，这显然比双向扩展的$(K-n-1)^2$个重复项减少了许多。

我们也可以从另外一个角度来分析。对于双向扩展而言，每个双字种子词能够扩展出$(K+1)^2$个多字项。而仅做向后的单向扩展时，每个双字种子词能够扩展出$(K+1)$个多字项，与前者相差了一个数量级。当然，两种策略中K的取值是不相同的：双向扩展向左右分别扩展K个字，得到的多字项最大长度为$2K+2$。为了获取同样长度的多字串，单向扩展必须向后扩展$2K$个字。因此，对于某个确定的K，采用双向扩展策略每个双字种子得到$(K+1)^2$个多字项，为达到同样的效果，单向扩展策略下每个双字种子得到的多字项为$(2K+1)$个。显然，在$K>0$的情况下，有

$$(K+1)^2-(2K+1)=K^2>0$$

可见，单向扩展会使得扩充而来的多字项大为减少。在一次对比实验中，采用单向扩展比采用双向扩展节约了 30.60%的时间，在系统效率上这无疑是一个比较大的进步。

当然，从理论上来说，单向操作的一个缺点在于：专业术语必须由双字种子向右扩展而得，如果某个术语的最开始两个字没能作为种子词汇抽取出来，那么就无法获得完整的专业术语。那么，实际效果如何呢？

在多次对比实验中，我们将用单向、双向扩展过程分别得到的术语候选项进行了比较，其重合率为均在95%以上。下面是不重合候选项的例子：

扩展方式	不重合项
单向	“大集团”、“诉记者”、“规划委”、“人民币”、“有限责任公司”、“延边市国有资”、“价值投资理念”、“次拍卖”、“韩志国”、“经营报”、“药业”、“治理”、“募集”、“一家”、“上报”、“市值”、“季度”、“条心”
双向	“利能力”、“有一家”、“年年初”、“公司上报”、“流通市值”、“告诉记者”、“监会规划委”、“三季度”、“业有限责任公”、“一条心”、“延边市国有”、“学家韩志国”、“中国经营报”、“监会主席”

表 4-4　单向与双向扩展不重合词汇举例

从上述实验可知，以最终结果而论，单向扩展与双向扩展相差无几。另外，还可以考虑降低阀值以获得更多的候选项。所以我们在 DSTES 系统的种子词扩展模块中选择进行单向扩展。

上面所说的第二个问题是：K 应该取什么值？

由第二章中的介绍可知，对于汉语而言，大多数词汇其长度都很短：最常用 9000 词，单字词、双字词、三字词占 99.2%[27]，超过三字的只占 0.8%；经过学者对专业术语的研究，其长度分布情况与通用词汇大致类似，也是长度小于某个固定阀值（该阀值与特定领域相关）的词汇占绝大多数。参考文献[21]中对 725 万字的信息领域专业文献中带英文注释的术语（汉英术语）进行了人工标记，然后利用程序提取汉英术语及其前界环境（前至少 4 个汉字）。在此基础上，获取了 2480 条全中文术语，其分布见下表：

字数	1	2	3	4	5	6	7	8	9	10	11	〉12
数量	20	337	275	594	296	405	198	147	79	57	28	31
比例（%）	0.81	13.59	11.09	23.95	11.94	16.33	7.98	5.93	3.19	2.30	1.13	1.77

表 4-5 信息领域术语分布表

因此，我们只需要在语料库 Corpus 中截取双字种子出现处前后的若干字，即可构成种子词汇所处的“环境”，用来获取完整的特定领域专业术语。当然，这个“环境”对于其余的计算语言学任务，如切分标注、句法分析等，是远远不够的。但是对于我们的专业领域术语抽取任务，已经是绰绰有余了。

综合前述研究成果，我们确定重点考察长度不超过 6 的多字串，即 $K = 4$。

第五章　系统测试结果及展望

为了验证本文提出的各种基于规则方法和基于统计方法在专业术语抽取领域的有效性，我们实现了特定领域专业术语抽取系统(DSTES)原型。以此原型为基础，进行了多方面的测试。本章首先介绍系统的测试语料，然后总结测试结果，最后根据测试结果提出了进一步完善方案。

5.1 测试语料

为了使测试结果能够很好的反映术语抽取系统的性能，用于测试的语料必须满足如下几个条件：

- 真实性：测试语料必须是从真实文本中选取的，不能是人工生造出来的；
- 广泛性：测试语料要尽量涵盖本领域内的各个方面，不能有失偏颇；
- 随机性：测试语料应该随机从语料库中抽取，以求覆盖更多的语言现象；
- 规模：除了有质的保证外，测试语料还需要有一定的量，才能更真实的反映系统的性能。

本系统采用的测试语料分别来自金融、足球和交通领域，下面的测试结果是在金融领域语料上进行实验得出的。这些金融语料主要来自金融网站，如搜狐财经（http://business.sohu.com）、新浪财经（http://finance.sina.com.cn）、中国金融网（http://zgjrw.com）等。由于这些语料大多数是新闻报道，所以有一定的时效性。前文已经介绍了一些方法来处理诸如“2002 年”之类的时间词汇，这里就不再赘述。

最后，我们在语料库中随机选取了 1756 篇金融语料（包含汉字 2，154，497 个）进行测试，测试平台为 CPU 1.6GHZ，内存 512MB。

5.2 测试结果

为讨论问题的方便，我们先做如下定义：

定义 5-1　MWU（Multi-Word Unit，多字单元）：具有完整意义的多字串，包括领域专业词汇和通用词汇以及它们的组合。

在具体介绍系统测试结果之前，有个问题必须首先明确：哪些抽取结果是“正确”的？

如前所述，在最终的结果中，有相当一部分词汇是 MWU，如“有一天”、“上海”等。尽管它们是有意义的词组，但并不是特定领域的术语，只对生语料切分等应用有一定作用，对于其他方面的应用，如专业术语库的建立、信息检索、

本体的建立等没有任何意义。

在以往的一些系统中，在统计系统的正确率、召回率时，对于抽取结果是否为术语并不予以区分，只要是有完整意义的词汇，就将其作为正确结果对待。这些系统抽取的其实是 MWU，而不是专业领域的术语。

为此，我们后面在介绍 DSTES 系统的测试结果时，给出两种结果：一个是针对 MWU 的结果 *PM*，以便跟其他系统的性能进行比较；另一个则是针对专业术语的结果 *PT*，这是体现我们系统特色的地方，即进行特定领域专业术语的抽取，而不是 MWU。*PM* 和 *PT* 的定义如下：

定义 5-2 $PM = C(M)/C(*)$ ，$PT = C(T)/C(*)$

其中，$C(M)$、$C(T)$、$C(*)$ 分别为抽取结果中 MWU 的个数、术语候选项的个数、抽取词汇总数。

5.2.1 测试结果

系统运行于上述测试语料，总共得到 2692 个词语，其中 MWU 有 2305 个，未能准确成词的有 387 个。在 MWU 中，有 350 个词汇为公司名称、人名、地名等专有词汇，按照参考文献[20]中的统计方法，我们得到 DSTES 系统的准确率 *PM* 为 72.62%，与参考文献[18]中的平均准确率 54.09%相比改善了许多。另外，由于 DSTES 系统的设计目标是抽取专业领域内的术语，而不是 MWU，因此在抽取过程中我们过滤掉了很多判定为一般词汇的 MWU，根据对中间结果的记录，这样的词汇有 859 个，所以，DSTES 系统的准确率与参考文献[20]74.4%的准确率相比，也有了一定程度的提高。下图是抽取得到的前若干个词的准确率：

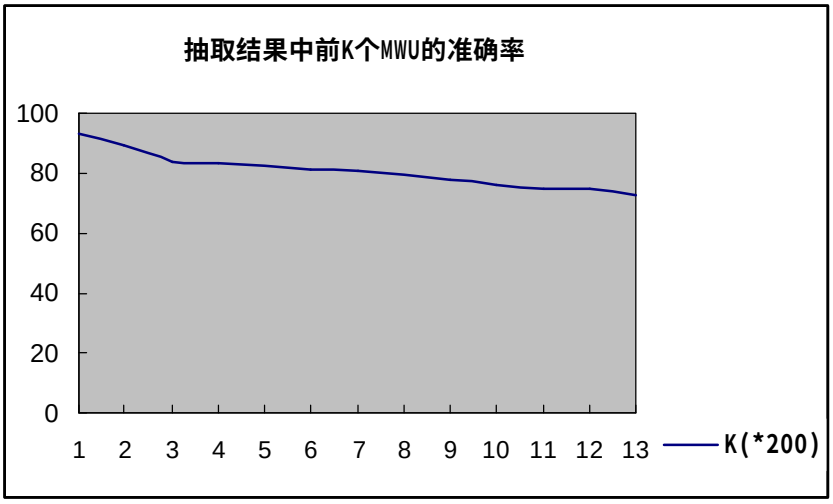

图 5-1　抽取结果中前 *K* 个 MWU 的准确率

经过筛选，我们在这 2692 个词语中，总共获得了 1381 个金融领域专业术语候选项，所以 DSTES 系统抽取术语的准确率 *PT* 为 51.30%。由于绝大多数现有系统给出的都是 MWU 抽取的准确率，而不是术语抽取的准确率，所以这里就不再进行比较。

5.2.2 系统其他性能

除了抽取的准确率，该系统的领域切换方便性也是我们所关注的。领域切换的方便性与系统的运行效率是一对矛盾：为方便的转换处理领域，系统中使用的方法必须尽量与当前处理领域无关，这显然会影响一些与领域相关、却能提高系统效率的方法的使用。

在领域切换方面，本系统也有不错的表现。从前文的介绍中可以看出，本系统与特定领域基本无关，只在过滤词表的一部分体现了一定的领域相关性。因此，如果要变换当前处理领域，只需要将过滤词表中与领域相关的部分进行相应的替换即可。我们做过从金融领域切换到足球领域的实验。结果证明，只需要变更过滤词表的小部分即可，非常方便。

5.3 工作展望

结合上面的测试与分析，本文最后提出进一步的工作展望。

第一，提高系统的运行效率。尽管我们采用了一系列的方法来提高系统的运行效率，但是由上面的测试结果可知，仍然不够理想，系统运行时间随语料规模的扩大增长很快。为了使系统能够有效处理更大规模的语料，必须进一步提高系统的效率；

第二，改进基于规则的方法。本文根据汉语词汇及术语本身的特点，提出了一些基于规则的过滤方法，取得了较好的效果。如果能进一步精确化这些处理规则，则术语抽取系统的性能将会进一步提高；

第三，扩大训练语料，优化统计模型中的参数。对于统计模型而言，其中的参数（即阀值）是至关重要的。因此，必须进行大量的训练以逼近最优参数。另外，还需要考虑训练语料过多时出现的过度训练问题。同时，还可以考虑针对特定领域进行训练、逼近最优参数，因为不同领域可能有不同的最优参数；

第四，尝试在线抽取专业术语。现有系统使用的语料是从网上下载到本机运行的，与网络资源相比，属于静态语料。如果能够直接处理网络上的大规模动态语料，那么我们便可以构建一个动态的术语抽取系统。这在信息爆炸的今天，无疑具有重要的意义。这种方式的主要难点之一

在于系统的运行效率，因为与本机相比，网络的存取速度相对较慢。另外，语料是否与领域相关也是值得研究的一个问题。

参考文献

[1] 中华人民共和国国家标准GB／T10112—959 术语工作 原则与方法（代替GB／T 10112—988）,1999；

[2] 冯志伟.现代术语学引论[M].北京.语文出版社:1997

[3] Sayori Shimohata, An empirical method for identifying and translating technical terminology. In Proceedings of the 17th conference on Computational linguistics, Volume 2, pp.782 – 788, 2000

[4] Kageura, Kyo & Bin Umino, Methods of Automatic Term Recognition: A Review. Terminology, 3(2): 259-289, 1996

[5] Luhn, H.P., A Statistical Approach to Mechanized Encoding and Searching of Literary Information. In IBM Journal of Research and Development 2(2): 159-165, 1957

[6] 郑家恒，杜永萍，宋礼鹏.农业病虫害词汇获取方法初探．第七届全国计算语言学联合学术会议论文集(JSCL-2003),语言计算与基于内容的文本处理，清华大学出版社

[7] Bourigault，D., Surface grammatical analysis for the extraction of terminological noun phrases. In Proceedings of COLING-92.pp.977-981.Nates, France, 1992

[8] Justeson, J.S. and Katz, S.M., Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. In Natural Language Engineering 1(1): 9-27, 1995

[9] Jones, L.P., Gassie, Jr., E.W. and Radhakrishnan, S, INDEX: The Statistical Basis for an Automatic Conceptual Phrase-Indexing System. In Journal of the American Society for Information Science 41(2): 87-97, 1990

[10] Kit, C, Reduction of Indexing Term Space for Phrase Based Information Retrieval. In Interim Memo of Computational Linguistics Program, Pittsburgh: Carnegie Mellon University, 1994

[11] Frantzi, K.T. and Ananiadou, S, Statistical Measures for Terminological Extraction. In Proceedings for the 3rd

International Conference on Statistical Analysis of Textual Data (JADT 1995):297-308, 1995

[12] Kita, K., Kato, Y., Omoto, T. And Yano, Y., A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria. In Journal of Natural Language Processing 1(1):21-33, 1994

[13] Frantzi, K.T., Ananiadou, S., and Tsujii, J, Extracting Terminological Expressions. In The Special Interest Group Notes of Information Processing Soceity of Japan, 96-NL-112, March 14-15 at Tokushima University, Tokushima, Japan, 1996

[14] Maynard, A. and Ananiadou, S, Identifying contextual information for multi-word term extraction. In Proceedings of Terminology and Knowledge Engineering Conference-99, pp.212-221, Innsbruck, Austria, 1999

[15] 陈文亮，朱靖波，姚天顺.基于Bootstrapping的领域词汇自动获取.第七届全国计算语言学联合学术会议论文集(JSCL-2003),语言计算与基于内容的文本处理，清华大学出版社

[16] Hiroshi Nakagawa, Hiroyuki Kojima and Akira Maeda, Chinese Term Extraction from Web Pages Based on Compound Term Productivity, In Proceedings of 3rd ACL SIGHAN Workshop, pp.79-85, 2003

[17] Smadja, F., Retrieving collocations from text: Xtract, Computational Linguistics, 19(1): 143 -177, 1993

[18] Fung, P., Extracting key terms from Chinese and Japanese texts. In The International Journal on Computer Processing of Oriental Language. Special Issue on Information Retrieval on Oriental Languages, pp.99 -121,1998

[19] Dogan, I. and Church, K., Termight: identifying and translating technical terminology. In Proceedings of Applied Language Processing, pp. 34-40, Stuttgart, Germany, 1994

[20] Patrick Pantel and Dekang Lin, A Statistical Corpus-Based Term Extractor. In: Stroulia, E. and Matwin, S. (Eds.) AI

2001, Lecture Notes in Artificial Intelligence, pp. 36- 46. Springer-Verlag, 2001

[21] 邢红兵.信息领域汉语术语的特征及其在语料中的分布规律，2001年10月26日，术语标准化及信息技术

[22] 龚益．规范社会科学术语势在必行，2003 年 02 期，社会科学管理与评论

[23] Bourigault, D, Lexter, a Natural Language Processing Tool for Terminology Extraction. In Proceedings of 7th EURALEX International Congress, 1996

[24] Church, K.W. and Hanks, P.P., Word association norms, mutual information and lexicography. In Proceedings of the 27th Annual Meeting of the ACL, pp.:76-83, Vancouver, 1989

[25] Dias, G., Guillore, S., Lopes, J.G.P., Mutual Expectation: a Measure for Multiword Lexical Unit Extraction. In Proceedings of VEXTAL Venezia per il Trattamento Automatico delle Lingue, 1999

[26] 北京语言学院语言教学研究所，现代汉语频率词典．1986：北京语言学院出版社

[27] Fung, P., Extracting key terms from Chinese and Japanese texts. The International Journal on Computer Processing of Oriental Language. Special Issue on Information Retrieval on Oriental Language, pp.99-121, 1998

[28] Liu, Y., New advances in computers and natural language processing in China. Information Science, 8:64-70, 1987

[29] Wu, Zimin and Gwyneth Tseng, Chinese text segmentation for text retrieval: Achievements and problems. Journal of The American Society for Information Science, 44(9):532-542, 1993

[30] Sproat, Richard, Chilin Shih, William Gale and Nancy Chang, A stochastic word segmentation algorithm for a Mandarin text-to-speech system. In Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics, 66-72,Las Cruces, New Mexico, 1994

[31] 罗盛芬．孙茂松，基于字串内部结合紧密度的汉语自动抽词实验研究．中文信息学报,2003 年 03 期

[32] Dunning, T., Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61-74, 1993

[33] 国家技术监督局，标点符号用法．1995 年 12 月 13 日

[34] 郭锐.现代汉语词类研究.北京：商务印书馆，2002

[35] 朱德熙.语法讲义.北京：商务印书馆，1998

[36] 北京大学中文系 1955、1957 级语言班.现代汉语虚词例释．北京：商务印书馆，1996

[37] 中国大百科全书·语言卷·术语.中国大百科全书出版社：2003

[38] R. Navigli and P. Velardi, Semantic Interpretation of Terminological Strings, In Proceedings of 4th Conference. Terminology and Knowledge Engineering(TKE 2002), 2002, Lecture Notes in Computer Science 2300, Springer-Verlag, New York, 2002, pp.325-353

致　谢

首先我要感谢我的导师陆汝占教授。陆老师对我给予了悉心的帮助与大量的指导。陆老师在数理逻辑与计算机科学及形式化模型论方面均有高深的造诣，并具有渊博的学识，对我的学术帮助很大。他的孜孜不倦追求的精神以及严谨的治学作风均对我产生很大影响，值得我们学习。

另外，陈晓明老师和裴炳镇老师也给予了我诸多的帮助与鼓励，在此向他们表示感谢。同时感谢陈玉泉副教授和高峰老师对我的指导。特别感谢吴蔚林、郭曙纶等对我论文方面的帮助，我还要感谢田怀凤、江丰、袁琰、王立、舒芳蕊等各位同学的帮助。同时感谢所有帮助过我的计算机系的老师和同学。

最后，我要感谢我的父母，感谢他们在我最艰难的时候给予我的支持。没有他们多年来的养育和教诲，没有他们在精神上和经济上的支持，我根本无法完成我的学业。

攻读硕士学位期间发表的论文

1. **杜波、田怀凤、陆汝占，基于多策略的专业领域术语抽取器的设计与实现，计算机工程及应用（已录用）**
2. **袁琰、杜波、田怀凤、陆汝占，基于框架的对话管理模型的研究与实现，计算机工程（已录用）**