中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：18722005

上海大学

硕士学位论文

SHANGHAI UNIVERSITY MASTER'S DISSERTATION

题目	基于工程测试技术本体的语义检索系统研究与实现

作　　者　博文
学科专业　机械工程
导　　师　胡小梅
完成日期　2021 年 6 月

姓　　名：博文　　学号：18722005
论文题目：基于工程测试技术本体的语义检索系统研究与实现

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主任：

委员：

导　　师：

答辩日期：　2021 年 6 月 21 日

姓　　名：博文　　学号：18722005
论文题目：基于工程测试技术本体的语义检索系统研究与实现

原 创 性 声 明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：博文　日 期：2021.6.21

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（**保密的论文在解密后应遵守此规定**）

签 名：博文　导师签名：[illegible]　日期：2021.6.21

上海大学工学硕士学位论文

基于工程测试技术本体的语义检索系统研究与实现

姓　　名：　博文

导　　师：　胡小梅

学科专业：机械工程

上海大学机电工程与自动化学院

2021 年 6 月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

Research and Implementation of Semantic Retrieval System Based on Engineering Testing Technology Ontology

MA Candidate：Wen Bo

Supervisor：Xiaomei Hu

Major：Mechanical Engineering

School of Mechanical & Electronic Engineering and Automation, Shanghai University

June, 2021

摘　要

随着工程测试技术的发展，此领域所包含的知识、技术以及经验都日益增多，对于此领域内的企业而言，将这些内容有效的进行知识管理以实现高效学习和技术传递就显得尤为重要。而现有的针对这类需求的知识管理工具不能满足用户的需求，如果能够提出一套针对此类需求的解决方案，来应用于工程测试技术领域，那么就能够减轻当前该领域对于企业级知识管理的需求的压力。

本文基于自然语言处理技术，结合领域本体，语义检索模型，本体构建算法等相关技术，提出了针对工程测试技术领域的知识管理应用开发的解决方案，并进行了基于工程测试技术本体的语义检索系统的实现。本文开展的研究内容如下：

首先，通过对现有的本体抽取算法的研究，先后提出了基于综合特征策略的术语抽取算法和基于主次聚类的术语关系抽取算法，二者共同组成了本体抽取算法。该算法克服了本体抽取算法过程中的不同特征带来的术语和术语关系的权重偏置问题，提出了针对不同特征的具体策略。经过实验，证实该算法有效提高了术语抽取和术语关系抽取的准确率。

其次，基于上述提出的本体自动抽取算法提出了基于综合特征策略的本体构建方法，并针对工程测试技术领域，进行了领域本体的构建，对本文提出的本体构建方法进行了实现，使用 protégé 工具生成了工程测试技术本体的 OWL 文件。

然后，通过分析当前信息检索的趋势，提出了用户需求模型，并基于用户需求模型和对经典信息检索模型的分析，提出了基于本体的需求融合信息检索模型，主要包含：查询代理、持久化、本体、经典模型、用户需求模型等五个模块。

最后，采用 java 语言进行了本体的 OWL 文件的解析与推理，完成基于工程测试技术本体的语义检索系统的设计与实现。主要包括：系统需求分析，系统介绍与平台选择，功能实现、系统测试与分析等。

关键词：本体抽取；工程测试技术本体；用户需求模型；语义检索

ABSTRACT

With the development of engineering testing technology, the knowledge, technology and experience contained in this field are increasing day by day. For enterprises in this field, it is particularly important to effectively carry out knowledge management on these contents to achieve efficient learning and technology transfer.However, the existing knowledge management tools for such requirements cannot meet the needs of users. If a set of solutions for such requirements can be put forward and applied to the field of engineering testing technology, the pressure on the demand for enterprise-level knowledge management in this field can be alleviated.

In this paper, based on natural language processing technology, combined with domain ontology, semantic retrieval model, ontology construction algorithm and other related technologies, a solution for the application development of knowledge management in the field of engineering testing technology is proposed, and a semantic retrieval system based on engineering testing technology ontology is implemented.The research contents of this paper are as follows:

Firstly, by studying the existing ontology extraction algorithms, a term extraction algorithm based on comprehensive feature strategy and a term relation extraction algorithm based on primary and secondary clustering are proposed successively, which constitute the ontology extraction algorithm together.This algorithm overcomes the problem of weight bias of terms and term relations caused by different features in ontology extraction algorithm, and proposes specific strategies for different features.Experimental results show that the algorithm improves the accuracy of term extraction and term relation extraction effectively.

Secondly, based on the ontology automatic extraction algorithm proposed above, an ontology construction method based on the comprehensive feature strategy is proposed. Aiming at the engineering testing technology field, the domain ontology construction is carried out. The proposed ontology construction method is implemented, and the OWL file of engineering testing technology ontology is generated by Protege

tool.

Then, through the analysis of the current trend of information retrieval, user requirement model is put forward, and based on user demand model and the analysis of classical information retrieval model, the demand of fusion information retrieval based on ontology model is put forward, mainly include: the query agent, persistence, ontology, classical model, five modules such as user requirement model.

Finally, the Java language is used to analyze and reason the OWL files of the ontology, and the design and implementation of the semantic retrieval system based on the engineering testing technology ontology are completed.It mainly includes: system requirement analysis, system introduction and platform selection, function realization, system testing and analysis, etc.

Keywords: Ontology extraction; Engineering testing technology ontology; User demand model; Semantic retrieval

目　录

第一章 绪论

1.1 课题目的和意义

随着互联网应用和信息交流方式的快速发展，信息化产品在为人们生活方式提供极大便利的同时，也使得数据呈指数级增长。据 IDC 报告显示，近几年全球的数据量以每年 58%的速度增长，未来数据产生的速度会更快。同时，国际调研机构 Forrest Research 的统计资料表明，只有 17%左右的结构化数据有效存储在各种类型的结构化数据库中，而其余 83%来自互联网的非结构化多语言文本数据分散在整个业务过程及外部环境中[1-5]。工程测试技术领域的数据量也随着时代的变迁而变得日益庞大，工程测试技术相关数据对测试的整个周期具备重要的意义。工程测试技术数据中最重要的数据类型之一就是文本数据，文本数据的知识管理与检索不可或缺。随着数据挖掘、机器学习、计算架构等技术的不断革新，文本检索模型的架构也逐步发生变化，使原本针对大多数群体的检索模型架构转变为更加注重于小群体或个人的检索模型架构[6-8]。

本文基于上述需求，设计并完成了基于工程测试技术本体的语义检索系统，该系统包含布尔检索以及语义检索功能，满足用户对于工程测试技术文本数据的知识管理需求，用户可基于此系统对企业或者组织的知识、技术或者经验进行管理和应用。并且提出了改进的术语抽取算法和术语关系抽取算法，二者共同组成了本体抽取算法，为本体构建提供算法基础。同时，针对语义检索的需求，提出了改进的语义检索模型。

开发基于工程测试技术本体的语义检索系统，不仅能够帮助对于文本数据的知识管理，而且对于其他领域的本体构建过程能够提供一定的帮助。对于提高知识管理效率而言，有效的知识组织方式将使得领域知识、技术以及经验得到最大程度、最高效率的利用。对于其他领域的本体构建而言，本课题提出的改进的本体抽取算法将加快本体构建效率，能有效的应对领域内容的更新迭代。此外，本课题通过提出改进的信息检索模型，提高了信息检索的准确率，因此，本课题的研究对于信息检索压力的缓解和文本语义检索也有十分重要的现实意义。

1.2 国内外研究现状

1.2.1 本体研究现状

本体论(Ontology)最早起源于哲学概念，其含义为事物的本源[9]。本体论被引入信息领域的契机是信息技术的不断发展而产生的需求。这个需求就是在九十年代初期，知识工程领域的研究过程中出现了两大瓶颈：知识重用和知识共享[10]。学者们希望可以将对知识的描述扩展到语义层面，并在此基础上建立起可由机器识别的新的知识模型，在此需求下，“本体论”概念被引入此领域。在此之后，针对本体的相关研究在各领域不断加深，成为目前研究的热点。本体分为通用本体、领域本体和应用本体。领域本体能够为相关领域提供通用的、明确的以及形式化的说明，为语义的实现打下了基础。目前，学者们在信息检索领域对于本体的应用增强了信息检索的智能性，它是语义检索领域的重要组成部分。

本体的构建是一个很复杂、要求严谨并且需要不断迭代的过程。在进行本体构建之前最重要的步骤就是选择合适的本体构建方法。合适的本体构建方法能够在实际构建过程中简化构建流程，节省大量时间，减少工程人员为之付出的时间成本和精力成本。目前主流的本体构建方法主要分为三类：手工构建方法、半自动化构建方法和自动化构建方法。

手工构建方法：在领域专家的辅助下，使用清晰的本体表述方式来构建本体，这种方式是一种纯手工的构建方式。手工构建方法的优点是准确性最高，能够按照人们的意愿完全的将本体构建出来。但是，该方法在实际应用过程中，完全按照本体构建人员的主观意愿来进行本体的构建，主观性太强，容易出现与实际应用所需要的本体存在较大差距的现象。并且，如果在信息量大的应用场景中，这种本体构建方法将会既耗费大量时间又耗费人力物力，无法快速构建，将对后续应用过程中的信息化过程、智能化应用过程带来不便甚至造成极大的困难，以至于所构建的本体存在共享性和重用性低等缺点，因此不适用于大型的本体构建。

半自动构建方法：半自动构建方法的一个重要特征是不完全依赖人工。这种方法可以复用已有的本体，使得本体构建的起点就是已有的本体。也可以使用计算机技术进行本体的构建，目的是缩短本体构建周期。但是会存在人工与自动构建分工不明确、无法找到现有本体以及本体合并过程的不统一等问题。

自动构建方法：自动构建方法将几乎完全避免人工参与，其以计算机技术为基础，以自然语言处理、机器学习与数据挖掘等技术为桥梁，以建立能够客观反映领域特点和逻辑关系为目的进行本体的构建。此类方法将通过对数据源的提取过程得到概念与概念之间的关系，以此为基础进行本体的构建。该方法的优点是能够不需要人工参与来进行本体的构建；缺点是数据噪声大，需要进行降噪工作的设计与实施，语义逻辑基础较为匮乏，概念之间的关系具备较低的可信度。表 1-1 展示了常见本体构建方法的特点。

表 1-1 常见本体构建方法特点

方法名	特点
TOVE	企业建模法，它是针对已有本体来设计的，为实现多种本体的集成而提供相应的解决方案和指导方法，可以集成和评估本体。
Skeletal Methodology METHONTOLOGY	骨架法，服务于本体开发时期，针对从本体构建初期的准备到最终的本体建立都提出了详细开发和构建的指导思想。总结了骨架法的指导思想，具有普遍的通用性。
Cyclic Acquisition Process	基于循环思想，由五个操作流程构成，依次是：数据源选择-概念学习-领域聚焦-关系学习-评价。
IDEF-5	利用结构化表述方法，使用图表以及具体的语言描述形式，来实现对知识中概念、属性以及他们之间关系的表达，并以此为指导思想构建本体。

1.2.2 术语及术语关系抽取算法研究现状

常用的术语及术语关系抽取方法大致分为三类：基于统计的方法、基于规则的方法、基于混合思想的方法[11]。这三类方法各有优缺点，不同的学者采用了不同的方法进行了术语的抽取。

基于统计的方法的基本思想是：以术语的统计规律作为依据，进行术语的抽取[12-15]。首先，要搜集大量领域相关的文本，以研究内容为依据来决定具体文本是否作为语料库进行术语及术语关系抽取。其次，对于候选术语进行基于文本的统计特征分析，将候选术语与领域的相关度与其在语料库中出现的频率关联起来，

各种统计特征的值越高，说明领域相关度越高，并通过设定阈值，将真正的术语从文本语料库中提取出来，形成术语集合。闫琪琪使用 C-MI 统计特征进行术语的自动抽取[16]，但召回率偏低。梁颖红提出了 NC-value 和互信息结合的方法进行术语抽取[17]，能有效识别长术语。刘辉、刘耀等人提出了基于条件场的术语抽取算法[18]。

基于规则的方法的基本思想是：通过对特定领域的术语的特异结构组成，分析其词法结构，在此基础上综合成术语的词法规则，进而构建词法模板，将之与语料库中的候选术语进行模式匹配，进而得出语料库的术语集合[19]。该类方法的术语提取准确度高，能够抽取文本中低频次的候选术语，相对基于统计的方法而言具有较强的词频无关性。俞琰等人提出了基于依存句法分析的术语抽取算法[20]。Azanzi Jiomekong 使用了克夫模型从 java 代码中提取本体[21]。Wen Zeng 提出了基于大量科技文献的术语抽取和相关分析[22]。

基于混合思想的方法的基本思想是：将基于统计的思想与基于规则的思想结合起来，进行术语的抽取。董洋溢、李伟华等人提出了将文本特征和统计量相结合的领域术语抽取方法[23]。刘里、肖迎元等人将术语长度结合语法特征进行了术语抽取[24]。

一、术语及术语关系抽取方法的优势与劣势对比

1、基于统计的方法：依靠统计方法的术语提取算法，对文本语料的要求如下：

（1）语料库所包含的文本数量足够大；

（2）语料库内的文本范围需要涉及到领域的各个方面。

要想达到较高的准确率，必须采用大量的文本语料，以达到文本语料库的全面性，这类方法一个特点是计算量大，最重要的是，文本语料库的“大小”的衡量尺度没有统一的标准[25]。难以保证能准确体现候选术语在领域的重要程度。基于统计思想的抽取算法的一个很重要的特征就是词频，词频是候选术语在语料库中的出现次数的归一化处理后的指标，而语料库是一个庞大的信息集合，其中会出现很多低频候选术语，对于这些词汇的领域相关性的判断是术语及术语关系抽取过程中很重要的一个部分，但是基于统计思想的方法并不能解决低频词汇的领域相关性的分析问题。另外，基于统计的术语抽取算法中有一类是基于机器学习

技术的。此类方法，常用的模型有最大熵模型、最大熵马尔科夫模型和条件随机场等，但是，即使采用了此类方法来训练学习模型，进一步的术语抽取时对于低频候选术语仍然不具备很好地效果。因此，单一的基于统计思想的术语抽取方法的准确率难以提升。

2、基于规则的方法：基于规则的术语及术语关系抽取算法，需要对文本中的术语的构成模式进行总结，进而以此模式为标准，进行术语的抽取。此类算法对设定的关键词构成模式依赖较大，术语提取的效果的优劣与规则的完善程度有关。并且难以用少量规则覆盖复杂的术语组成规律，并且当规则数达到一定数量时会产生相互冲突问题。因此，基于规则的术语抽取算法的术语识别效率有限。

3、基于混合思想的术语及术语关系抽取算法：此类算法融合了规则和统计理论的思想，进行术语抽取算法的改进。研究此类算法的专家学者将根据自己的侧重点对算法进行改进，总体来说能够改善单一的基于统计或者基于规则的方法的术语抽取效果。

二、TF-IDF 方法介绍

TF-IDF 方法是基于统计学理论的术语抽取方法，以 TF 值和 IDF 值的高低来识别某候选词是否为领域术语。其中，TF 值是指词频，是指候选词 i 在语料库中的文档 a 中出现的频率。如果该候选词汇在文档中出现的频率较高，则认为该候选词汇能够很好的代表此领域。但是此单一的文档特征的局限性太大，并非每个领域文本中的频率最高的候选词汇都会是领域术语，因此，学者们使用 IDF 特征来进行误差的平衡[26-27]。IDF 是逆向文件频率，此统计特征是由总文件数目 D 除以包含该词语的文件数 d，再将得到的商取以 10 为底的对数得到。

三、C-value 方法介绍

C-value 方法是一种基于语言学规则和统计理论的混合术语抽取方法，以 C-value 值的高低来识别术语[28]。C-value 值与其在领域语料中的词频成正比，与其长度成正比，与被嵌套的程度成反比。C-value 方法简单、适用性强，具有语言和领域无关性，在长术语和嵌套术语的抽取方面极具优势。

1.2.3 信息检索模型研究现状

一、信息检索模型概述

信息检索是用户进行信息查询和获取的主要方式，是查找信息的方法和手段。狭义的信息检索仅指信息查询，即用户根据需要采用一定的方法，借助检索工具，从信息集合中找出所需要信息的查找过程。广义的信息检索是将信息按一定的方式进行加工、整理、组织并存储起来，再根据用户特定的需要将相关信息准确的查找出来的过程，又称信息的存储与检索[29-30]。一般情况下，信息检索指的就是广义的信息检索。每个信息检索系统都遵循一定的检索模型，并且一个信息检索系统的核心就是信息检索模型。经过学者们的不断研究与积累，目前的信息检索模型可分为三类：布尔模型、向量空间模型、概率模型。

二、布尔模型

布尔模型的特点是：简单和严格匹配。这种检索模型以集合理论和布尔代数作为其基本理论[31]。

布尔模型的核心思想是：索引词在文本中只有两种情况，即出现与不出现。因此，布尔模型的权值变量都是由二值（0,1）数据组成。在布尔模型中，用户查询将转换为析取范式，该析取范式所包含的简单合取式即是不同的满足查询的布尔表达式。

布尔模型的主要优点是显而易见的：形式简洁、结构简单。但是其不足之处也相当明显：针对一词多义或者一义多词的情况会导致返回的数据集合过多或过少[32-33]。

布尔模型的优点主要体现在一是运算速度快，二是简单易行可操作性强。但它也有很多缺点，主要体现在：

（1）利用布尔模型的检索结果非常不易控制。对于一个用户特定的查询，它可能检索到许多数据，参考性不强，也可能什么也检索不到，这是由于匹配条件要求过于严格，这导致了布尔检索的漏检率较高。

（2）利用布尔模型的检索只能反映信息的相关与否，而对于相关性大小的比较则不能进行，更不能反映特征项在文献中的重要程度。

（3）利用布尔模型的检索不能识别功能词，比如“有关…”这类检索指令，而这对于习惯了自然语言检索的用户造成很多不便。

三、 向量空间模型

布尔模型为向量空间模型的诞生提供了基础的思想，基于布尔模型的二元权

值的特性，向量空间模型提出了新的检索逻辑。

向量空间模型将用户的查询要求和数据库文档信息用检索项的向量空间来表示，其查询结果根据向量空间的相似性进行排列[34]。有效的查询结果在向量空间模型中可方便地产生，它也能提供相关文档的文摘，还可以对查询结果进行分类，为用户提供的信息更加准确。

向量空间模型的核心思想是：以向量来表示文本$(W_1, W_2, W_3, \ldots W_n)$，其中$W_i$表示第 i 个特征项的权重。而对于一个文本的向量表示的基本步骤就是：首先，对文本进行分词，将这些词作为向量的维数表示；其次，以权重的形式表示该词的出现与否；最后，持久化到数据库中。对于权重的确定方式而言，W_i一般通过词频、文本数量等依据来确定每个特征项的权值。

向量空间模型相对于布尔模型的主要优点是：克服了布尔模型二元权值的缺点，采用非二元权值来表示特征项在文本和用户查询中的权重，提出了允许部分匹配的模型结构。

四、概率模型

给定一个用户的查询串，返回一个包含所有与该串相关的文档的集合。这是一个理想的结果文档集，通过这个文档集，我们会很容易得到结果文档。它的具体处理过程是：第一步，用户大致浏览一下结果文档，在这些文档中区分出相关的与不相关的；第二步，系统利用用户给予的信息重新定义出比较理想的结果集的概率描述；第三步，再重复以上两步操作。这样做就会越来越接近所需要的结果文档集。

概率模型理论是假设用户查询串与文档相关的概率只决定于查询串和文档。换句话说，该模型假定这样一种情况，存在着这样一个集合，它包含了所有文档，恰恰它又与查询串的结果文档子集完全相同，这种理想的集合用 R 表示，那么这个理想的集合中的文档一定是与查询串相关的[35]。

概率模型的优点是：其一，采用严格的数学理论为依据，为人们提供了一种数学理论基础来进行检索决策；其二，采用相关反馈原理，可开发出理论上更为坚实的检索方法。它的主要缺点也很明显：一是增加存储资源和计算资源的开销；二是参数估计难度较大。

1.3 论文的主要研究内容

本文主要对基于工程测试技术本体的语义检索系统进行相关理论和方法的研究，具体的内容包括以下几个方面：

第一章 绪论。介绍了课题的目的，意义，以及相关理论技术的国内外研究现状。主要包括本体、术语（关系）抽取以及信息检索模型的研究现状。

第二章 术语及术语关系抽取算法优化。工程测试技术领域知识体系复杂，本体抽取算法是进行本体构建的核心环节，只有先将文本中的术语以及术语关系先抽取出来，才能进行本体的合成。本文将本体抽取算法分两部分处理：基于综合特征策略的术语抽取算法和基于主次聚类的术语关系抽取算法。基于综合特征策略的术语抽取算法对于术语抽取过程中产生的不同特征进行了分析，针对不同的特征提出了具体的策略。基于主次聚类的术语关系抽取算法主要使用二次 K-means 聚类结合特征提取的方式抽取术语关系，同时对 K-means 聚类算法进行改进，通过实验验证了两个算法的优越性。

第三章 工程测试技术领域本体的构建。首先通过分析领域本体的构建原则以及工程测试技术领域的特点，提出了基于综合特征策略的本体构建方法。然后，按照本文提出的本体构建方法的步骤，一步步对工程测试技术领域本体进行构建，最终得到了工程测试技术领域本体。最后设计了该领域本体的持久化方法，并生成了该本体的 OWL 文件。

第四章 基于本体的需求融合语义检索模型的设计。本章首先通过分析目前信息检索的趋势，提出了用户需求模型。然后，提出了基于本体的需求融合语义检索模型，主要包括以下五个模块：查询代理、持久化、本体、经典模型、用户需求模型。最后给出了不同模块的主要内容以及他们之间的逻辑关系。

第五章 基于工程测试技术本体的语义检索系统研究与实现。首先，介绍了系统的软硬件平台和要求。其次，分析系统的需求功能，详细说明了本体推理和数据库等核心技术，对第三章生成的的 OWL 文件进行解析，探究了信息检索系统的开发技术。然后，基于 java 语言完成系统的开发，并对主要界面进行展示。最后，进行了系统测试与分析。

第六章 结论与展望。

论文组织框架如图 1.1 所示

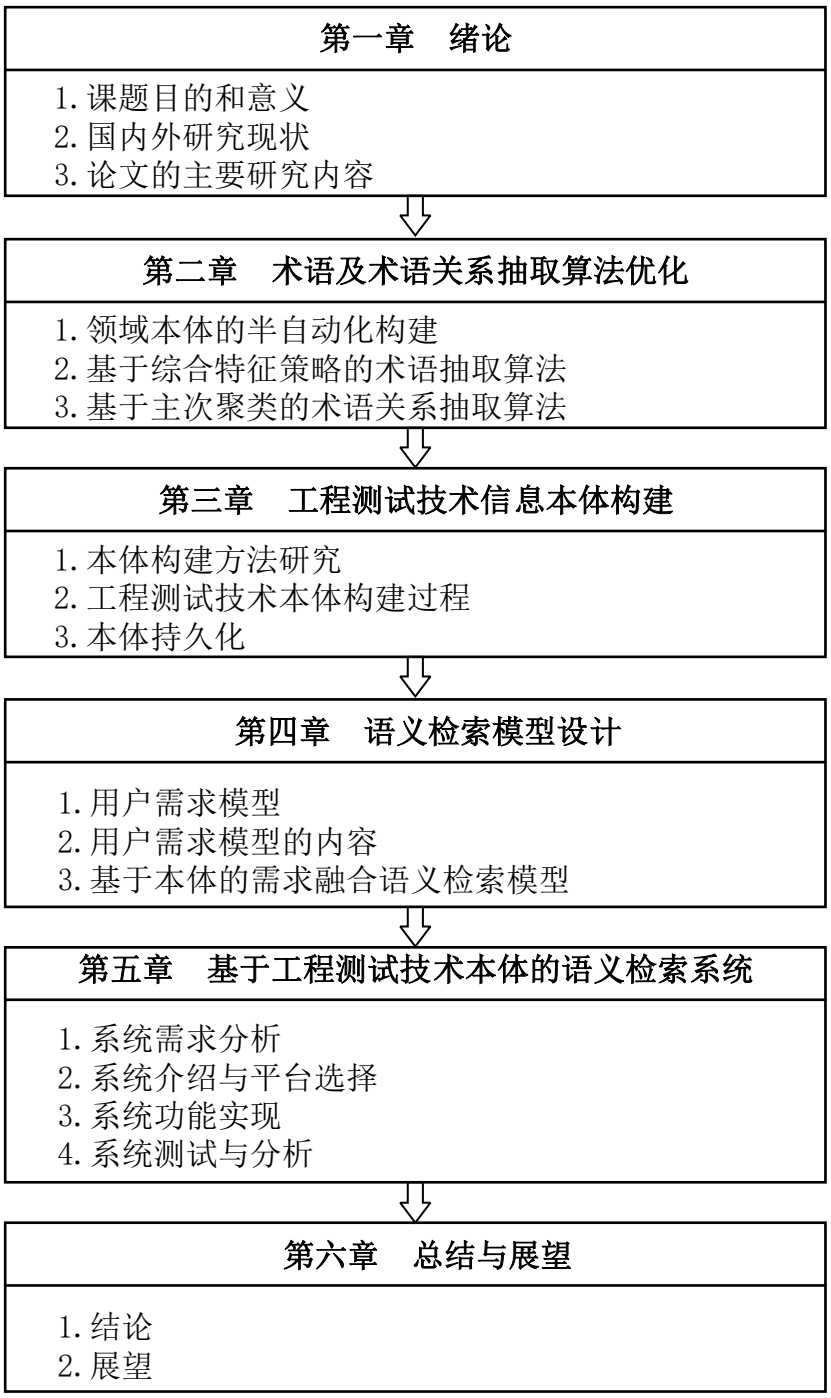

图 1.1 论文组织框架

第二章 术语及术语关系抽取算法优化

术语抽取和术语关系抽取是整个本体构建过程中极其重要的一步。本章将针对目前的术语抽取和术语关系抽取算法的不足，对术语抽取和术语关系抽取算法进行优化，提出基于综合特征策略的术语抽取算法和基于主次聚类的术语关系抽取算法，二者共同组成本体抽取算法。

2.1 领域本体的半自动化构建

本文提出的是基于综合特征策略的本体构建方法，是一种半自动本体构建方法。领域本体的半自动化构建过程中能够体现自动化的主要过程是术语抽取和术语关系抽取两个过程[36]，二者的核心内容和他们之间的联系如图 2.1 所示：

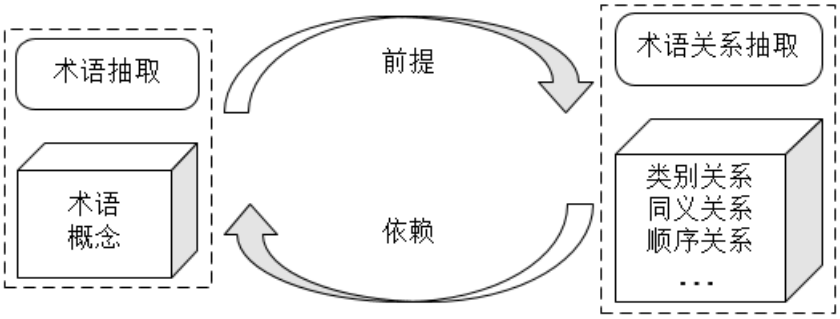

图 2.1 本体自动化构建的核心内容

术语抽取是本体半自动化构建的第一步，它的核心是在语料库中抽取领域需要的或者企业应用需求指定的领域术语。所有的术语共同构成了本体知识库这个知识网络中的每一个节点。例如，机器人设计与制造领域中存在的典型术语：“机械臂”、“机械手”、“机器视觉”、“控制流程”、“路径规划”、“自由度”等。

术语关系抽取是本体半自动化构建的第二步，它的核心是在语料库中抽取领域需要的或者企业应用需求指定的领域术语关系。所有的术语关系共同构成了本体知识库这个知识网络中的每一条路径[37-39]。例如，农业机器人领域中存在的典型术语关系：“控制了”、“反馈”、“采摘”、“传输了”等。

本文设计的本体构建方法中：术语关系抽取模块将在术语抽取模块之后进行，并且将术语抽取的结果作为术语关系抽取的基础。术语抽取是进行术语关系抽取的前提，术语关系抽取的效果的优劣将依赖于术语抽取结果的准确率，二者共同

承担了领域本体抽取的主要任务。

2.2 基于综合特征策略的术语抽取算法

2.2.1 基于综合特征策略的术语抽取算法概述

基于目前的术语抽取算法的研究现状，其中基于混合思想的算法能够取得更好的效果，而一般学者们提出的基于混合思想的算法中考虑的语言学特征较为单一，容易造成准确率下降[40-44]。因此，为全面适应领域的术语抽取工作，本文提出一种基于综合特征策略的术语抽取算法，以期能解决中文术语抽取算法的稳定性，以此提高术语抽取的准确率。图 2.2 展示的是本文的术语抽取算法的思想路线。

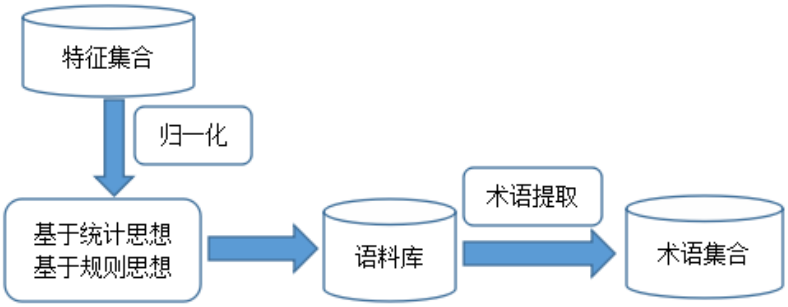

图 2.2 术语抽取路线

2.2.2 基于综合特征策略的术语抽取算法

一、术语抽取过程的综合特征

本文将术语的特征分为两个方面：未经分词的术语特征、经分词的术语特征[58]。术语的抽取首先要将文本进行切分，将非结构化文本转换为候选术语的集合。本文将候选术语在生语料中就已经具有的特征定义为未经分词的术语特征。而由于中文分词产生的新术语特征定义为经分词的术语特征。

1、未经分词的术语特征

未经分词的术语特征是由文本语料本身决定的。由于术语是一个领域内关键概念的体现，是描述性文本中最核心的词汇以及高频词汇，其重要性不言而喻。而在描述性文本其作用便是阐述领域的核心概念和理论。

（1）强调性特征：一般学者或者技术人员在通过文本描述自身领域的相关

理论和概念时，将会着重突出本领域的重要概念，而约定俗成的会将关键概念和理论在特定位置进行强调，本文称之为强调性特征。领域文本中，术语会有更大的可能性出现在完整文本的首段和末段，那么，在进行术语提取时就有必要将该特征进行数学上的体现以提高术语抽取效率。另外，文本内部的每一段落的首句也是术语出现的最佳可能位置，同样。在进行术语抽取时应当突出该位置特征。因此，本文将提出一个公式用于提升候选术语的权值，用以突出相关候选术语的术语可能性的值的大小。

（2）TF-IDF 特征：依据现有的研究成果，在未经分词的文本语料库中，术语的 TF-IDF 值也是领域文本中术语的关键特征。

（3）偏置度特征：本文对偏置度特征的定义如下：偏置度是候选术语只出现在文本的某一部分的强度。之所以将偏置度作为一个单独的特征，是因为在使用术语抽取算法的过程中，候选术语的偏置度越大，说明该候选术语是领域术语的可能性越大。因此本文将之作为术语抽取的一个重要依据。

2、 经分词的术语特征

经分词的术语特征是在文本转换成候选术语集合后的新增的特征。中文的术语抽取研究不同于其他语种的术语抽取，需要先进行分词，再根据分词的结果进行下一步研究。本文定义两种特征：过度切分特征和误切分特征，本文将之统称为经分词的术语特征。

（1）过度切分特征：本文将文本分词过程中产生的将一个真正的领域术语切分成两个或者两个以上候选术语的现象称之为过度切分特征。分词产生的候选术语中非领域术语也占据了很大一部分的原因就是由于文本在进行切分时产生了过度切分的现象，将原本是一个术语的候选词汇切分成了几个候选词汇。

（2）误切分特征：这是经分词的术语特征的另外一个特征，是术语的误切分问题产生的。误切分特征是指文本分词过程中产生的将一个领域术语添加了多余字符的现象。例如，将形如“计算机科学”的候选术语切分成“计算机科学与”类似的词串，这也是候选词切分过程中产生的一个重要问题。

二、术语抽取的综合策略

基于上述对于术语抽取过程中的特征分析，本文针对于不同的特征带来的问题进行了综合策略的研究。综合策略包含两个部分：基于未分词的术语特征的术

语抽取策略和基于经分词的术语特征的术语抽取策略。对于策略的综合描述如下表 2-1 所示：

表 2-1 特征与策略关系映射表

分词类型	特征类型	策略
未分词	强调性特征	对不同强调性术语进行权重增强。
未分词	偏置度特征	对不同偏置度术语进行权重增强
未分词	TF-IDF 特征	使用逆文档算法。
经分词	过度切分特征	定义语义扩展法来尽可能还原术语形态。
经分词	误切分特征	定义语义扩展法来尽可能还原术语形态。

1、未分词术语特征抽取策略

本文提出的未分词术语抽取策略是针对未分词术语特征的特点设计的术语抽取方法。上述未分词类型下的术语特征包含三个方面：强调性特征、偏置度特征和 TF-IDF 特征。本文针对未分词特征的抽取策略是，在传统的 TF-IDF 算法的基础上，加入基于候选术语的强调性特征和偏置度特征的抽取策略，三者相结合，得出一种改进的 TF-IDF 计算公式，对术语的抽取算法进行改进。

基于 TF-IDF 特征，本文采用传统的 TF-IDF 公式，对于 TF-IDF 值高的候选术语的权值加以提升，对于该值低的术语的权值加以降低，以突出具有该特征的候选术语的权值。

基于候选术语的强调性特征，本文提出的抽取策略如下：在语料库处理时得出术语的强调性特征，并对候选术语的领域术语可能性的权值进行设定。当候选术语出现在领域文本的首末两段时，赋予该候选术语权值为 2；当候选术语出现在领域文本的段落首行时，赋予该候选术语权值为 1.5；当候选术语出现在领域文本的除上述两个位置的其他位置时，赋予该候选术语权值为 1。以下是本文设计的公式：

$$Ch=\begin{cases}2 & 出现在首末两段\\ 1.5 & 出现在段落首行\\ 1 & 出现在一般位置\end{cases} \quad (2.1)$$

公式 2.1 中，Ch即表示不同位置的候选术语的权值。

基于偏置度特征，本文提出的抽取策略如下：基于对偏置度的计算，对于偏

置度高的术语。提升其权值，对于偏置度低的术语，降低其权值。以下是本文设计的偏置度计算公式：

$$Intens = \frac{d_i}{D} \times \sum_{i=0}^{n} \frac{c_i}{C_i} \tag{2.2}$$

其中，$Intens$表示词汇偏置度，D表示经过过滤的候选词汇出现的次数总和，d_i表示在所有候选词汇中词汇 i 出现的次数，C_i表示出现了词汇 i 的所有段落中的候选词总和，c_i是指每个段落中出现的词汇 i 的频次，n表示出现词汇 i 的段落数。

本文基于以上的特征和策略，提出候选术语的权值计算公式如下：

$$TC = TF - IDF \cdot 0.5(Intens + Ch) \tag{2.3}$$

$$TC = \sum_{j=1}^{n} \frac{n_{i,j}}{\sum_k n_{k,j}} \cdot log \frac{|D|}{\{j: t_i \in d_j\}} \cdot 0.5 \cdot \left(\frac{d_i}{D} \times \sum_{i=0}^{n} \frac{c_i}{C_i} + Ch \right) \tag{2.4}$$

式中， $n_{i,j}$表示候选词 i 在文档 j 中出现的次数，$n_{k,j}$表示候选词 k 在文档 j 中出现的次数。D 表示语料库的文档数，$\{j: t_i \in d_j\}$表示包含候选词的文档数。公式中其他的符号含义上文已经详细解释，在此不再赘述。

2、经分词的术语抽取策略

经过文本分词之后，候选术语出现了过度切分以及误切分的特征，并且此类候选词汇一般较长（4 字以上）。根据此特点以及 C-value 方法在抽取嵌套词的优势，本文采用语境扩展法和 C-value 方法相结合的策略来进行术语的提取。

对于过度切分的候选术语，本文定义一种语境扩展法将该类候选术语进行扩展。首先，根据文本预处理时的分隔符，得到候选术语之间的位置关系。其次，由于过度切分的候选术语将产生更少的字数，一般将少于 8，于是将术语的字数范围定义为 2~8，即以语料库段落为范围，将候选术语向左侧以及右侧进行术语扩展，得到新的候选术语的组合。最后，使用 C-value 方法进行术语的提取。

误切分的候选词汇往往同时属于低频词汇，因此，可进行语料库的全面处理。此时，语境扩展法改进为：将该低频词汇进行相同字数以及上下波动两字的候选术语进行选择，形成误切分术语的集合。候选词扩展后进行 C-value 值的计算，根据 C-value 值大小进行排序，设定阈值，筛选出大于阈值的候选术语。

三、术语抽取的特征策略模型

基于上述研究，本文提出术语抽取的特征策略模型如图 2.3 所示：

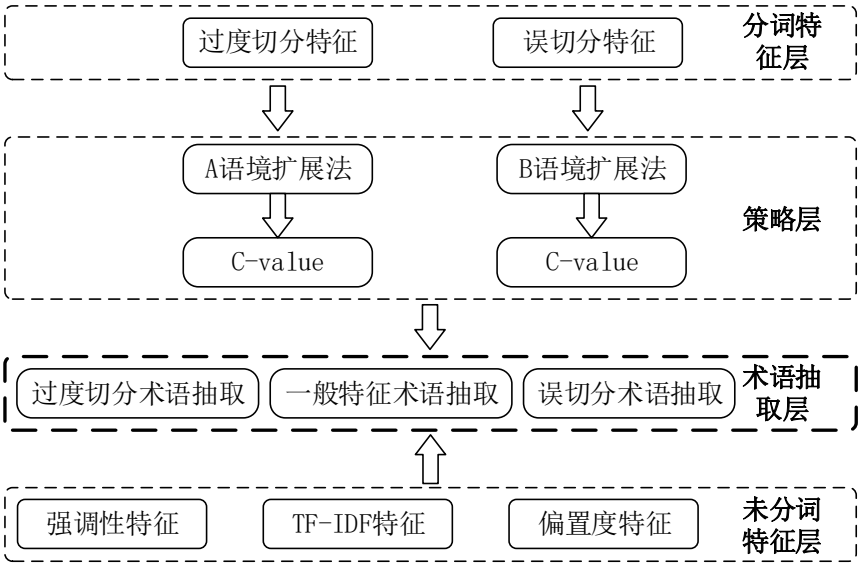

图 2.3 特征策略模型

该特征策略模型是基于上述的对术语抽取过程的特征以及对应的策略详细设计后提出的关系模型。主要分为四层：分词特征层、策略层、术语抽取层以及未分词特征层。分词特征层完成权重增强后进入术语抽取层，分词特征层在经过策略层后也进入术语抽取层，共同完成术语抽取的任务。

四、基于综合特征策略的术语抽取算法流程

在进行术语抽取算法描述之前，先介绍文本预处理的方法：

（1）选取领域文档和一般文档作为术语提取的语料库和对比语料库。对比语料库的作用是将不具有领域特征的词汇进行剔除，包括：停用词、通用词以及无实际意义的词汇；

（2）使用分词系统分别对两类文档进行分词处理得到词汇集合，设定过滤阈值 th，在领域文档中除去一般文档词频超过 th 的词汇，得到候选词集合 D。

在通过文本预处理得到算法要求的输入格式之后，就可以开始术语抽取的流程。通过对以上的特征和策略的分析与研究，本文最终提出了基于综合特征策略的术语抽取算法。该算法的基本思路是：通过对领域术语特征的综合分析研究，对术语特征加以整合，采用分层次分策略的细粒度提取方法进行领域术语的抽取。对于此基于综合特征策略的术语抽取算法的要求是：1、能够有效提高术语抽取的准确率；2、能够具备较好的通用性，在不同的领域术语抽取过程中都能够有

相同水平的表现。以下是本文算法的流程图：

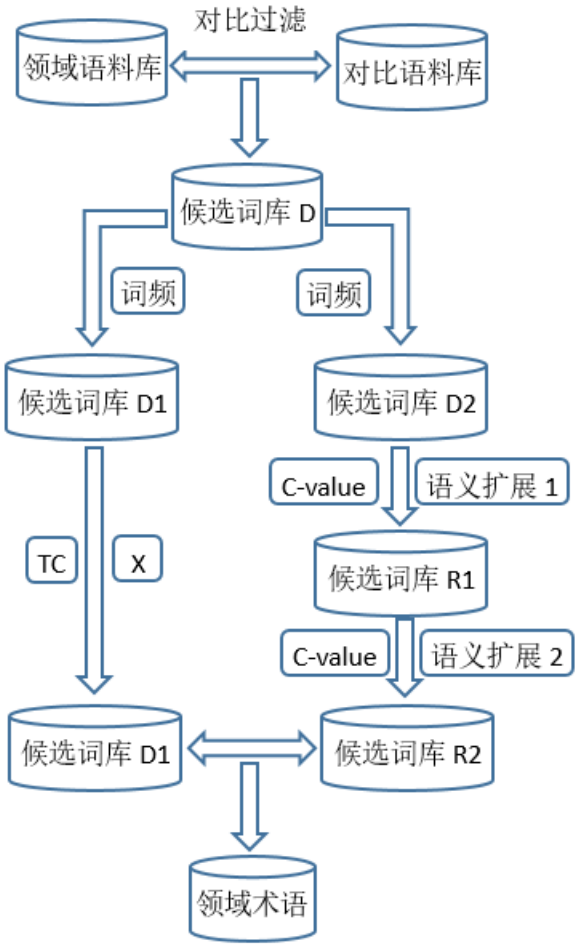

图 2.4 基于综合特征策略的术语抽取算法

下面是对基于综合特征策略的术语抽取算法的描述：

step1：将文本预处理后得到的候选词集合 D 分为两部分，字数小于等于 4 的候选词属于集合 D1，字数大于 4 的术语集合 D2；

step2：对于 D1 集合中的候选词，使用本文提出的改进的 TF-IDF 公式 TC 来计算其值，并进行排序，并设定术语抽取阈值 x，大于阈值 x 的存入集合 result；

step3：对于 D2 集合中的候选词，实行误切分特征对应的策略。将其中的候选词按照上文提出的语境扩展法进行候选词扩展，并计算 C-value 值，设定阈值 y，将超过阈值 y 的候选词添加到集合 R1 中；

step4：对于 D2 中的候选词，实行过度切分特征对应的策略，将其中的候选词按照上文提出的语境扩展法进行候选词扩展，并计算 C-value 值，设定阈值 z，将超过阈值 z 的候选词添加到集合 R2 中；

step5：将 R1 和 R2 集合中的术语进行合并过滤，若 R1 中的术语是 R2 中的

术语的子串，则将其删除；

step6：将经过 step5 处理的 R1 和 R2 集合中的术语存入到 result 集合中，术语抽取结束。

2.2.3 术语抽取实验与验证

一、领域语料库与对比语料库

本文的语料库获取方法是通过从相关网页上下载的方式获取到不同领域的文档将其转换为 txt 格式的文本语料库。从典型机械零件设计技术的领域中挑选 100 篇文本文档作为领域语料库，从其他工程制造领域挑选 600 篇文本文档作为对比语料库。

二、术语抽取效果评价方法

一般对术语抽取效果的评价指标采用准确率（P）和召回率（R）两个指标，为了综合判断术语的抽取效果，本文采用 F-score（F 值）指标来进行评价。计算公式如下：

$$F-score=\frac{2PR}{P+R} \tag{2.5}$$

三、术语抽取实验

本文术语提取实验过程如下：

（1）本文的语料库的获取是从维基百科抽取到的典型机械零件设计技术的领域网页语料库，抽取文档数为 100 篇，以相同方式获取其他工科设计技术领域文档 600 篇，将之转换为 txt 文件格式；

（2）使用中文分词系统 ICTCLAS 进行分词处理，并进行词性标注。将自然语言表达的文本文档转换成候选术语的集合，示例片段如下图所示：

机械/n设计/vn是/vshi机械/n工程/n的/n重要/a组成部分/nl，/wd是/vshi机械/n生产/vn的/n第一/m步/qv，/wd是/vshi决定/v机械/n性能/n的/n最/d主要/b的/n因素/n。/wj机械/n设计的/vn的/n努力/an目标/n是/vshi：/wm在/p各种/rz限定/v的/n条件/n（/wkz如/v材料/n、/wn加工/vn能力/n、/wn理论/n知识/n和/cc计算/vn手段/n等/udeng）/wky下/f设计/v出/vf最/d好/n的/n机械/n，/wd即/d做出/v优化/v设计/vn。/wj

图 2.5 中文分词与标注片段

（3）根据对比语料库，将领域语料库中与对比语料库中出现的高频重合词汇进行剔除，词频设定为 10，并去除无意义的虚词，数词等；

（4）根据所得文本语料库，进行人工标记，得出 483 个术语，形成术语集合 A；

（5）使用本文的基于综合特征的术语抽取算法进行术语抽取，其中阈值 x 设定为 4.0，y 设定为 2.0，z 设置为 3.0。得出术语集合 B。

（6）将所得术语集合 B 与集合 A 进行比较，得出准确率，召回率和 F 值；

（7）分别使用 TF-IDF 方法与 C-value 方法进行术语抽取，并分别计算其准确率、召回率和 F 值。

四、术语抽取结果分析

本文算法是基于规则和统计相融合的术语抽取算法，因此，实验结果的对比对象包括：基于规则的术语抽取算法和基于统计的术语抽取算法，由于统计的术语抽取算法中 TF-IDF 算法使用最广泛，因此，选择其作为对照，而基于规则的术语抽取算法中 C-value 方法较为经典，并且本文也对其进行了应用，因此，将其作为另外一个对比的对象。同时，与同样引用两种经典术语抽取算法的基于似然比与 C-value 结合的术语抽取算法进行对比[45]。通过与以上三种算法的对比，验证本文算法的优越性。表 2-2 展示了三种算法的各项评价指标的对比：

表 2-2 算法效果对比

术语抽取方法	准确率（P）	召回率（R）	F-score
TF-IDF	87.7%	62.8%	73.1%
C-value	48.2%	64.7%	55.2%
基于似然比与 C-value 相结合的术语抽取算法	85.8%	74.5%	79.8%
本文基于综合特征策略的算法	94.7%	81.4%	87.5%

实验结果分析：

从表 2-2 可以看出，在中文分词系统进行分词处理的基础上，基于综合特征策略的术语抽取算法在准确率和召回率上，较两种传统算法和基于信息熵的术语抽取算法均有提升。F-score 较 C-value 方法和 TF-IDF 方法提升了 14.4%，较基于信息熵的术语抽取算法提升了 7.7%，很好的改善了术语抽取的效果。因此，本文的算法在术语抽取方面取得了明显的改进。

2.3 基于主次聚类的术语关系抽取算法

针对术语关系抽取，本节采用聚类算法自动获取术语关系。将术语的词法特征、语法特征及语义特征等综合起来，基于此获取术语间的各种关系，包括：同义关系、并列关系、属种关系、整体部分关系。本节首先提出了基于特征的变密度 K-means 聚类算法，然后在此基础上阐述了基于主次聚类的术语关系抽取算法，最后，通过实验进行了算法效果验证。

2.3.1 领域本体中术语关系的分类

用于构建本体的关系抽取与普通的关系抽取任务的不同之处在于：一方面要尽可能的保证关系的准确，另一方面要尽量减少关系抽取的代价。传统的关系抽取任务往往不需要考虑什么样的关系应当用于本体构建，往往也不需要考虑均衡人力代价和结果的准确率。本文通过分析领域本体中术语关系的分类，确定了重点研究的关系类型。

术语关系能够表征术语的组成部分或术语间是如何联系的。目前研究者从不同角度对本体中的术语关系进行分类。根据术语间的相似性，可以把术语关系分为同一关系、属种关系、交叉关系、全异关系、否定关系等五种关系[46-47]。根据代表术语的个体在空间或时间上的连接性可分为：空间上的整体部分关系，如引擎是汽车的一部分；时间上的连续关系，如在生物体的发育过程中，蚕会从幼虫生长为蛹。根据国家标准，术语间关系分为层级和非层级关系。层级关系包括属种关系和整体部分关系，同一层级间则为并列关系；非层级关系有序列关系及联想关系，序列关系即空间、时间、因果、源流及发展关系，联想关系有推理、形式-内容、函数、物体-属性、结构-功能、行为-动机、行为-客体、生产者-产品及工具-操作关系等。目前，像通用本体 WordNet, HowNet 等，都包含了层级关系和同义关系。在术语关系的研究中，选取了 4 个具有代表性的关系：同义关系，并列关系，整体部分关系和属种关系。其中整体部分关系和属种关系是层级关系，是构建本体所必须的；同义关系是构建概念的重要依据；并列关系是关系过滤和关系扩展的重要依据。这四类关系是领域本体构建任务中最重要的四类关系。下图是术语关系的核心分类：

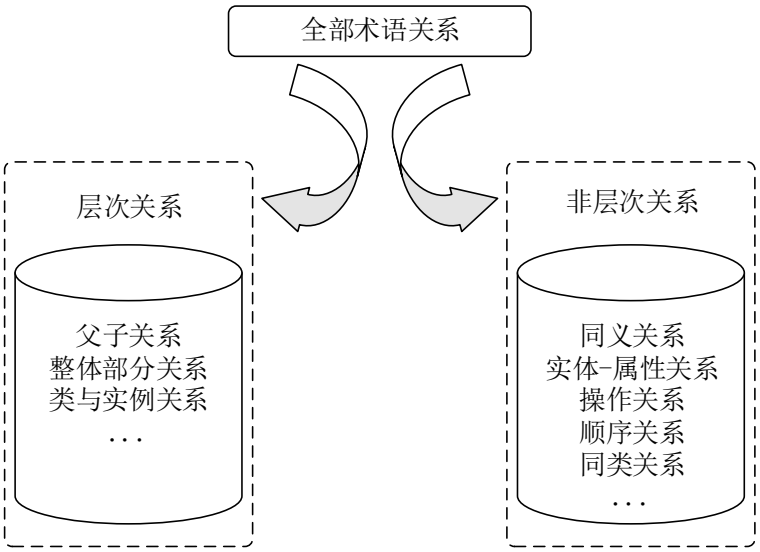

图 2.6 核心术语关系分类

2.3.2 基于主次聚类的术语关系抽取算法

针对术语关系抽取过程，考虑到有监督的机器学习算法需要大量的标注语料，其结果对于标注集有较大的依赖，而且标注语料的代价十分昂贵，因此这里结合聚类算法进行术语关系抽取。

本文通过总结规则来获取并列关系和同义关系；通过依存分析从语法结构的角度选择特征，使用聚类算法过滤掉不包含关系的实例；根据整体部分关系和属种关系的差异性再次选择特征和进行聚类，最终获得整体部分关系和属种关系。一种关系可以用三元组的形式表示，即(术语 1，关系，术语 2)。在要抽取的 4 类关系中，同义关系将用于同义术语合并进而形成概念。并列关系在实际问题中常常作为一个中间结果，起到过滤和扩展的作用。属种关系和整体部分关系则确定本体的骨架。主要包括以下几个核心内容：

（1） 文本数据的预处理；

（2） 核心关系；

（3） 规则匹配；

（4） 依存结构；

（5） 候选术语关系对；

（6） 特征向量；

（7） 基于统计的关系过滤；

（8） 基于特征的变密度 K-means 聚类算法。

基于主次聚类的术语抽取算法的框架如图 2.7 所示：

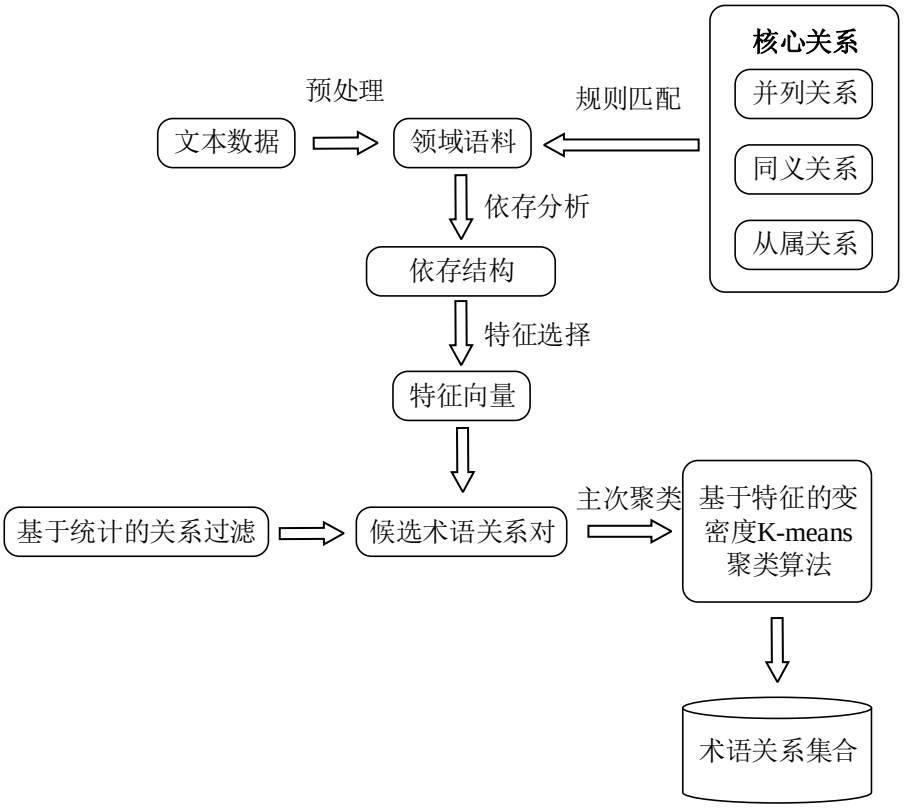

图 2.7 基于主次聚类的术语关系抽取算法框架

一、文本数据的预处理

文本数据的预处理主要包含两个部分：上下文环境语料和术语抽取过程产生的领域术语集合。上下文环境语料的处理方式类似于术语抽取过程的文本处理方法，在此不再赘述。将术语按照文本出现的顺序以及段落结构进行整合，以为算法的数据输入做准备。

二、核心关系

由于每个术语都有可能是其他术语的一个子类也可能是一个父类，所以本文将从属关系作为规则匹配的一个指标，用于快速筛选与其他术语不具备关系的术语。类似的核心关系还有并列关系和同义关系。

三、规则匹配

规则匹配是术语之间关系抽取的关键步骤，本文将给出规则匹配的具体方法：首先，在已经预处理的语料中采用随机抽样的方式，对术语关系对进行人工分析总结，得出大多数同义关系和并列关系都符合的某种规则。其次，进行规则的总

结归纳，最后，通过已经得出的关系类型对领域语料进行标记，后续直接作为术语关系对存入关系集合。

四、依存结构

为了得到句子的依存结构，本文对每一个句子进行依存解析，提取实例的深层句法结构特征来作为聚类的前提。本文采用哈尔滨工业大学 LTP 平台进行依存分析，在处理较长句子和带有省略成分的句子时，LTP 平台能自动的完成解析工作。

五、候选术语关系对

在术语抽取完成之后，对在一个句子中共现过的术语进行两两组合，每两个术语组成一个候选术语关系对，因此他们可能存在某种关系。

六、特征向量

本步骤将确定一个候选术语关系对的向量空间，比如，两个术语之间的动词本身、两个术语之间的动词在文档中的位置、两个术语之间、两个术语共现的频率、两个术语自身的 tf-idf 值等，都可以作为一个候选术语关系对的向量维度。

七、基于统计的关系过滤

具有紧密联系的术语对往往会在同一句子中共同出现，反言之，在文档中没有经常共现的术语对往往没有紧密联系。基于这个假设，按照如下公式对术语对进行过滤:

$$sim(t_1, t_2) = \frac{2 \cdot df(t_1, t_2)}{df(t_1) + df(t_2)} \tag{2.6}$$

其中$df(t_1, t_2)$表示术语t_1, t_2共同出现的句子数，$df(t_i)$表示术语t_i出现的句子数。该公式通过基于术语共现的规律得到了两个术语之间的相似度。

八、基于特征的变密度 K-means 聚类算法

K-means 聚类又称作 K 均值聚类，该聚类算法广泛应用于数据挖掘的各个领域[48-52]。其大体思想为：先任意选择初始中心点，其次，每次迭代过程中计算每个样本点到每个中心点的距离，计算公式如下：

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_i - x_j)^2} \tag{2.7}$$

公式 2.7 中，$D(x_i, x_j)$表示欧氏距离，x_i和x_j表示数据点,k 表示向量的某一维

度，m 表示向量的总维度。

并将样本划入距离该样本最近的中心点所在的类，然后重新计算中心点继续迭代，直到每个类不再发生变化为止。本文将两次使用 K-means 聚类，通过这种主次复合聚类的方式来实现对术语关系的分类。

K-means 聚类算法是一种无标签算法，算法效率比有监督算法要高，但是有三点明显的劣势：第一，任意选择的初始中心点可能会导致聚类不可控；第二，聚类中心 K 的取值的任意性不能保证聚类的准确性；第三，点与点之间的距离算法不能满足不同位置特征的差异。通过对以上三点的研究，本文提出了改进的方法：基于特征的变密度 K-means 聚类算法，改进的主要方面如下所述：

第一，对初始点选择方法进行改进。首先，遍历一次数据集，对每个点计算其密度大小，使用一个排序数组按密度大小的顺序记录每个数据点，遍历完成后，得到密度最大的数据点，将该点作为第一个初始聚类中心，使用欧式距离将其附近的点划分到该类中。第二个聚类中心的选择就是密度仅次于第一个聚类中心的数据点，以此类推，可根据 K 值来选择相应的聚类中心，K 值的确定方式将在第三点进行阐述。

第二， 欧式距离公式的改进。由于传统欧式距离的计算没有考虑到数据点的不同特征对距离的影响权重，因此本文对此加以改进。本文将术语抽取算法中提出的强调性特征与偏置度特征共同作为特征权重对欧氏距离公式加以改进，公式如下所示：

$$D_h(x_i, x_j) = \sqrt{\sum_{k=1}^{m} 0.5(Intens + Ch)(x_i - x_j)^2} \qquad (2.8)$$

公式 2.8 中，$D_h(x_i, x_j)$表示改进的加权欧氏距离，x_i和x_j表示数据点,k 表示向量的某一维度，m 表示向量的总维度。Intens 和 Ch 分别表示某一向量维度的强调性特征和偏置度特征，二者是本文在术语抽取算法中提出的，在此不再赘述。

第三，聚类中心个数 K 的确定方式的改进。合适的 K 值将使得类内的数据点距离小、类间的距离大。对于类内的数据点的距离计算，本文提出如下公式：

$$D_a = \left[\sum_{i=1}^{p}\sum_{j=1}^{n} 0.5(Intens + Ch)(x_j - m_i)^2\right]/n \qquad (2.9)$$

公式 2.9 中，D_a表示类内数据点到聚类中心点距离的平均值，i 表示聚类中心点的下标，p 表示聚类中心点个数，m_i表示聚类中心点，n 表示类内数据点个数，Intens 和 Ch 分别表示某一向量维度的强调性特征和偏置度特征。

对于类间距离的计算，首先计算不同聚类中心点的距离，然后选择其中的最小值来表示类间距离：

$$D_b = min(m_i, m_j) \tag{2.10}$$

公式 2.10 中，D_b表示类间距离的最小值，m_i和m_j表示不同的聚类中心点。

最终，本文定义的优化函数为：

$$M(p) = \frac{D_b - D_a}{D_b + D_a} \tag{2.11}$$

公式 2.11 中，$M(p)$即为不同 K 值取值时的优化函数，D_a和D_b分别表示类内平均距离和类间最小距离。由此可知，函数的取值范围为$[-1,1]$，$M(p)$越接近于 1，类内差异相对于类间的差异就基本可以忽略不计，此时代表聚类效果越好；$M(p)$越接近于-1，则刚好相反，说明聚类效果越差。因此，为了聚类效果最优，$M(p)$取最大值时，对应的 p 值即为最佳聚类中心的数目 K。

以上就是本文提出的改进的聚类算法：基于特征的变密度 K-means 聚类算法，并且，在对 K-means 聚类算法改进的基础上，为了进一步提高聚类效果，将使用主次聚类的方式进行处理，将每个实例看作是一个样本，两次聚类分别采用不同的特征。主次聚类包括主聚类和次聚类，第一次聚类的目的是去掉不存在关系的术语对。首先选择术语名称、术语前词词性、句子长度、依存路径长度、特殊符号特征，再利用 K-means 算法将存在整体部分关系或属种关系的实例聚为一类，将不存在关系的实例剔除。经过第一次聚类，过滤掉了大多数不存在关系的术语对，此为主聚类。第二次聚类的目的是将成对的术语按照不同的关系进行分类。选取了与第一次聚类不同的特征集合：依存路径长度、词袋特征和中间词特征。这些特征更能表征一个实例中所涉及的两个术语是整体部分关系还是属种关系，此为次聚类。

2.3.3 基于主次聚类的术语关系抽取算法实验与验证

一、语料特点及预处理

语料来自百度搜索“齿轮制造流程”返回的相关文本。经过去除标签、符号标准化、长短句切分等预处理之后，总共得到 11036 个句子。6127 个术语总共出现了 21560 次。经过统计，在 6127 个术语当中，有 3214 个术语仅仅出现过一次，认为这部分术语过于稀疏，应当去除以免影响术语关系的抽取准确度。

术语关系候选词对的获取按照以下方法来进行：首先，确定剩余的合格术语，本文将上述预处理过程中获得的术语集合去除只出现一次的术语；其次，对文本集合进行遍历，获得出现候选术语共现现象的句子；最后，将包含两个及以上候选术语的句子作为存在术语关系的句子，而这些对应的候选术语就组成了术语关系候选词对。本实验最终总共形成了 34509 个候选术语关系对。下一步从这些候选术语关系对中区分哪些术语关系对包含关系，以及包含什么样的关系。

二、规则分析

本文采用随机抽样的方式，对 34509 个术语关系对中的 1550 个进行人工分析总结。发现大多数同义关系和并列关系都符合某种规则。比如，图 2.8 中所列举的句子中的术语经总结后发现同义关系的术语满足一下三种规则的一种：

（1）规则 1：term；

（2）规则 2：term（同义关系词 term）；

（3）规则 3：term 同义关系词 term。

其中同义关系词包括：“叫做”、“又称”、“也叫做”、“也称为”、“还称为”、“即”、“就是”、“即是”。使用规则匹配的方法能有效的提取出同义关系和并列关系。并列关系则需要满足以下规则：

（1）规则 1：term 并列关系词 term；

（2）规则 2：term 并列关系词 term；

其中并列关系词包括：“和”、“与”、“还有”、“以及”。

例句 1：…采用圆弧渐缩齿（又称格利森制）或摆线等高齿（又称奥利康制）。
例句 2：如滚刀铲磨机床、多功能剃齿刀磨床、螺旋锥齿轮刀具磨床等。
例句 3：按零件结构可分为盘齿和轴齿…。
例句 4：如用于圆柱齿轮加工的滚刀、剃齿刀、插齿刀，用于直齿锥齿轮的圆拉刀，用于螺旋锥齿轮加工的各种铣齿刀具。

图 2.8 规则分析的句子举例

三、关系过滤

在一个句子中同时出现的术语之间往往具有紧密的联系，并且如果术语之间同时出现的概率越高那么他们具有紧密联系的概率就更大；反之，两个术语没有在同一个句子中同时出现或者同时出现的频率越小，那么他们之间往往没有紧密联系。

四、依存解析

本文对每一个句子进行依存解析，并通过提取实例的深层句法结构特征来完成聚类。本文基于 LTP 平台将基于 XML 的形式来对句子中的各个组成部分之间的依存关系进行展示。句子的依存解析结果可以表示成一棵树，树中的每个节点都是句子中的组成成分，包括标点符号和词，树中的边指示出两个节点之间的依存关系，依存树的示例如图 2.9 所示：

```
<sent id="0" cont="圆柱齿轮和锥齿轮的加工需要采用不同的工序组合。">
    <word id="0" cont="圆柱" pos="n" ne="undefined" parent="1" relate="ATT" semparent="5" semrelate="Agt"/>
    <word id="1" cont="齿轮" pos="n" ne="undefined" parent="5" relate="ATT" semparent="3" semrelate="eCoo"/>
    <word id="2" cont="和" pos="c" ne="undefined" parent="3" relate="LAD" semparent="3" semrelate="mConj"/>
    <word id="3" cont="锥齿轮" pos="n" ne="undefined" parent="1" relate="COO" semparent="5" semrelate="Feat"/>
    <word id="4" cont="的" pos="u" ne="undefined" parent="1" relate="RAD" semparent="3" semrelate="mAux"/>
    <word id="5" cont="加工" pos="v" ne="undefined" parent="7" relate="SBV" semparent="-1" semrelate="Root"/>
    <word id="6" cont="需要" pos="v" ne="undefined" parent="7" relate="ADV" semparent="5" semrelate="ePurp"/>
    <word id="7" cont="采用" pos="v" ne="undefined" parent="-1" relate="HED" semparent="6" semrelate="dCont"/>
    <word id="8" cont="不同" pos="a" ne="undefined" parent="10" relate="ATT" semparent="10" semrelate="Feat"/>
    <word id="9" cont="的" pos="u" ne="undefined" parent="8" relate="RAD" semparent="8" semrelate="mAux"/>
    <word id="10" cont="工序" pos="n" ne="undefined" parent="7" relate="VOB" semparent="11" semrelate="Agt"/>
    <word id="11" cont="组合" pos="v" ne="undefined" parent="7" relate="VOB" semparent="7" semrelate="dCont"/>
    <word id="12" cont="。" pos="wp" ne="undefined" parent="7" relate="WP" semparent="5" semrelate="mPunc"/>
</sent>
```

图 2.9 依存树示例

五、术语关系特征选取

本文将选取 6 种特征用于关系聚类。本文的聚类方法是：基于聚类过滤的方式来进行术语关系的聚类。首先，第一次聚类时将取三个特征：术语、术语词性、句子长度、依存路径长度；第二次则选取：依存路径长度和中间词特征。

六、聚类

使用本文提出的基于特征的变密度 K-means 聚类算法进行聚类工作。

经过聚类之后最终得到的结果如表 2-3 所示：

表 2-3 术语关系抽取的实验结果

关系类型	并列	同义	类属	整体与部分	综合
K-means 准确率	87.6%	60.7%	72.4%	64.5%	70.2%
主次聚类算法准确率	96.5%	70.4%	71.6%	74.8%	85.3%

从上表可以看出：本文提出的改进的术语关系抽取算法是有效的，相对于传统 K-means 方法在准确率上提高了 15.1%。

2.4 本章小结

本章通过对现有算法的研究提出了改进的术语抽取和术语关系抽取算法，二者是本体抽取算法的核心。首先，对领域本体的自动化构建进行分析研究，得出在构建领域本体时的主要内容；其次，进行术语抽取的算法研究并提出了基于综合特征策略的术语抽取算法；然后，进行术语关系抽取的算法研究并提出了基于主次聚类的的术语关系抽取算法；最后，通过实验来验证了术语抽取算法以及术语关系抽取算法的优越性，基于综合特征策略的术语抽取算法相对于其他基于混合思想的术语抽取算法 F 值提高了 7.7%，基于主次聚类的术语关系抽取算法相对于传统 K-means 方法在准确率上提高了 15.1%。

第三章 工程测试技术领域本体构建

基于第二章提出的本体抽取算法，本章进行了工程测试技术领域本体的构建。工程测试技术是实验科学的一部分，主要研究各种物理量的测量原理和测量信号分析处理方法。工程测试技术体系中不仅包含有大量的测量及试验相关知识，还包括一些具有工程测试价值的工程经验，传统的数据组织方式便捷快速，但对于隐含的内在属性及关系无法有效提取。尤其是当知识到达应用层级时，知识的共享、检索等外部接口或应用需要有更加模块化、层次化的知识组织方式。工程测试技术本体为这种知识组织问题带来了解决和探索的方向，本章基于本体理论，以工程测试技术为领域最终构建工程测试技术领域本体，提取出核心概念、关系等关键信息，提高了知识的利用层次，也避免了不同知识源中相同概念产生的语义异构问题。

3.1 本体构建方法研究

3.1.1 本体构建原则

目前业内已经构建了许多应用于不同领域的本体，这些本体的构建者、企业或组织基于领域特点，从应用背景和用途等不同角度出发，提出了各自的本体构建方法，这些建模方法都普遍遵从了 Grube 在定义本体时所给出的 5 项原则[53]：

（1）Clarity（清晰性）：对于相关内容范围内的术语语义，本体应该用准确的、清晰的语言去描述和定义。

（2）Coherence（一致性）：本体中包含的知识和含义应该是一致的，不能产生相冲突的地方，即便是经由语义推理而产生的知识或概念，也应该是和本体内原有的该知识或概念相同。

（3）Extendibility（可扩展性）：本体应该是可扩展的，它应该支持在原有的知识上，引入新的知识，而无需对本体中其他内容进行修改。

（4）Minimal encoding bias（编码偏好程度最小）：对于本体的表述形式，不能对某种特定的编码层产生依赖，而是要使用通用的、大家广为认可的符号表述形式，来对本体进行构建。

（5）Minimal ontological Commitment（本体约定最小）：对于本体概念和知识的定义，我们应该只给出最少的约束即可，因此在构造本体的过程当中。我们通常需要邀请领域专家来进行帮助。

3.1.2 基于综合特征策略的本体构建方法概述

基于以上原则和对现有本体构建方法的研究，本文提出了一种改进的本体构建方法：基于综合特征策略的本体构建方法。本方法是一种本体半自动构建方法，其中，基于综合特征策略的术语抽取算法和基于主次聚类的术语关系抽取算法已经在第二章详细介绍。基于综合特征策略的本体构建方法的内容如图 3.1 所示：

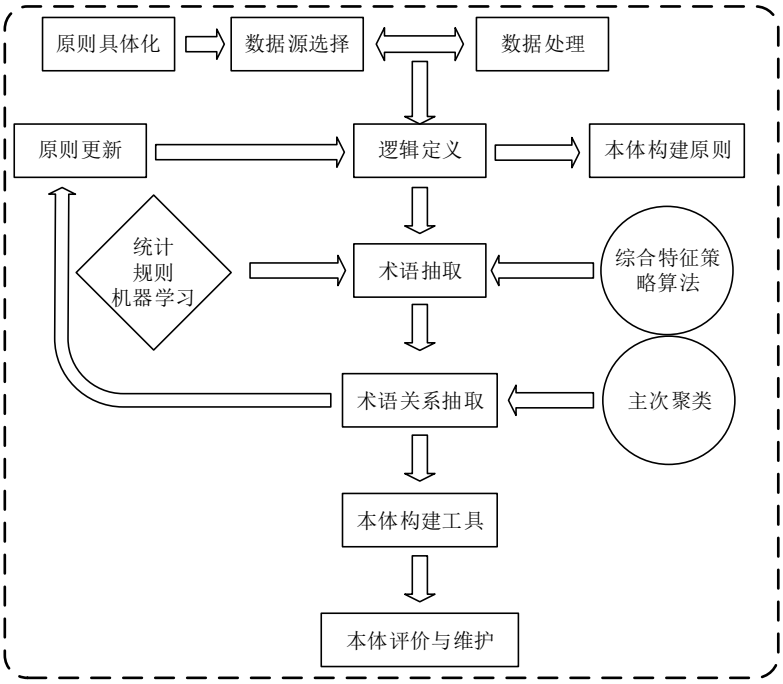

图 3.1 基于综合特征策略的本体自动构建方法

3.1.3 基于综合特征策略的本体构建方法详细设计

一、原则具体化

在本体构建的过程中，只有抽象的原则作为指导显然是不够的。因此，本文将本体构建原则进行了具体化分析。主要分为两个方面：自动化程度分析和可操作性研究。

（1）自动化程度分析

不同的应用场景或者应用需求对自动化程度的要求的不一致性，在进行本体

构建时并不一定要采用完全自动化的方式。例如，一个检索系统中存在两个不同的模块 a 和 b，a 模块对检索的需求是语义检索，b 模块对于检索的需求是关键词检索。那么，在进行本体构建时只需要对 a 模块的概念进行自动化构建，而 b 模块则可以将全部关键词作为检索的起点，不需要对其进行自动化的构建，人工定义将更加合适[54-56]。对具体领域本体的构建的自动化程度分析的意义在于，工程人员可以依据该过程的分析结果选择合适的对本体的构建方式。

（2）可操作性研究

可操作性研究主要内容和目的是分析得出在某个领域建立领域本体的可行性。本文提出的可操作性研究主要包括以下几个指标：概念的存在性、语料的丰富度、数据筛选的标准以及应用的现实意义。

二、数据源选择

本文提出的数据源选择方法的主要内容包括：数据领域范围的确定、数据大小的限定以及数据获取方式的选择。第一，数据领域范围的确定。在进行领域本体的构建时，并非将一个领域的所有内容全部包含进来，而是通过对应用需求的分析来对所需要的数据范围加以确定。第二，数据大小的限定。在确定了数据领域范围后开始收集领域数据时，要依据计算机的处理能力以及算法的效率选择适当的数据大小。第三，数据获取方式的选择。本文的数据源主要来自于非结构化文本。在进行术语抽取和术语关系抽取之前，首先需要构建语料库。文本数据的获取方法有：（1）使用别人做好的语料库；（2）可利用爬虫等技术获取数据并构建自己的语料库。

三、数据处理

数据处理的目的是对上一步数据选择的结果进行处理，将无效数据、脏数据以及不符合行业特点的偶然数据进行剔除，并作为下一阶段本体自动抽取算法的输入[57-59]。主要内容包括：文本选择性剔除、停用词剔除、中文分词、数据整合、格式化数据包。

文本选择性剔除是指当原始数据中的非文本数据。停用词剔除是将文本中的类似于“的”、“呢”等类似的词剔除的一个过程，一般是对照停用词表进行剔除。中文分词就是将字符串按照一定的规范重新组合成词序列的过程。数据整合是将已经进行过中文分词的数据进行统计分类的过程。将不同候选术语出现的次数、位置等信息进行统计。为后续术语抽取和术语关系抽取做基础。最后，为了能够

作为算法的输入，还要格式化数据包，将数据的格式调整成为算法设计时对算法输入要求的格式。

四、逻辑定义

本模块的主要任务是对领域数据的主要逻辑关系进行分析，比如，父子关系，并列关系，因果关系等。主要内容包括：术语类别统计、术语关系统计以及形成逻辑集合等三个部分。

第一，术语类别统计。领域本体的主体部分便是术语集合，因此，本文提出了对术语类别统计的具体方法：首先，将经过数据收集步骤和数据处理步骤之后的语料文本作为输入，进入一个正则表达式统计术语类别的算法；其次，将每个术语的频率组成结果集；最后，将同一类别的术语组成同一个集合。这样，就可以获得本体中术语的类别组成。

第二，术语关系统计。领域本体的另外一个核心部分便是术语关系集合，不同术语关系类别的比重将影响本体应用中技术实现的方式。本文提出了对术语关系类别统计的具体方法：首先，将经过术语类别统计的术语集合语料文本作为输入，进入一个正则表达式统计术语关系类别的算法；其次，将每组术语之间的关系的频率组成结果集；最后，将同一类别的术语组成同一个集合。并且，加入实际应用中需要特别强调的术语关系集合。这样，就可以获得本体中术语关系的类别组成。

第三，形成逻辑集合。在经过上述三个步骤之后，逻辑集合便可以确定。本步骤的主要任务是将术语之间的关系进行形式化的记录和持久化的保存，并且，在后续对本体进行修正时，本模块应当提供本体更新的接口。

五、术语抽取

领域本体构建的一个主体部分就是术语抽取[60]。本文的术语抽取过程使用第二章提出的基于综合特征策略的术语抽取算法，算法的具体介绍在此不再赘述。

六、术语关系抽取

领域本体构建的另外一个主体部分就是术语关系抽取[61]。首先，术语关系抽取本身就需要具备较高的准确率；其次，术语关系抽取效果的优劣还将影响到下一步本体应用的效果。如果术语关系抽取过程出现严重问题，那么本体的应用效果也会变差。本文的术语关系抽取过程使用第二章提出的基于主次聚类的术语关系抽取算法，该算法已经在第二章详细介绍，此处不再赘述。

七、原则更新

在上述几个模块中，构建本体所需要的一些指标或者参数已经在经验或者实验分析中满足，但是本体在进入一个实际的应用场景之后就会出现需要更新或者重新修正本体的需求。原则更新模块将通过信息检索模型中的检索反馈模块，将在实际应用过程中，不断更新信息检索的频率特点，更新本体构建的原则。原则更新主要包括以下几部分内容：

（1）术语内容的更新。在实际应用场景中，针对企业的领域本体会随着行业或者企业的发展而变化，即有些术语将不再使用或者新增一部分术语。那么，本体构建者就应当及时进行术语内容的更新[62]。而术语更新的节点可以由企业以应用中实际的效果进行决策是否更新。

（2）术语关系的更新。在牵涉到具体应用时，针对企业的领域本体中的术语关系比重会随着行业或者企业的发展而变化，即有些术语关系将不再使用或者新增一部分术语关系[63]。那么，本体构建者就应当及时进行术语关系的更新。而术语关系更新的节点可以由企业以应用中实际的效果进行决策是否更新。

（3）应用需求的更新。由上文所述可知，应用需求将导致本体构建原则的变化，因此，在进行原则更新时，应当考虑对本体应用时的应用需求的变化。应用需求的不同将带来术语比例以及术语关系比例的不同。

八、本体的构建过程分析

根据本文提出的基于综合特征策略的本体构建方法，本文通过对本体的构建过程进行分析，分别得出一个重要的指标，为本体的评价标准作基础。

（1）对文本的相关度进行分析。这一过程处于数据的收集和处理阶段，这个阶段决定了术语抽取和术语关系抽取效果的上限。

（2）对术语抽取的准确度进行分析。术语抽取将产生一个术语集合，本步骤的任务只考虑抽取出的术语在全部术语中的准确度。如果某一部分的术语明显缺失，则对上一步文本相关度的设定加以调整。

（3）术语抽取和术语关系抽取之间的关系分析。由于术语关系抽取是在术语抽取的基础上进行的，因此，在分析术语的抽取效果时必然要依据术语抽取的侧重点。例如，术语抽取时侧重动名词或者名词的抽取，则术语关系抽取过程的分析也应当侧重于动名词的词对或者名词的词对来进行分析。

（4）术语关系抽取效果分析。术语关系抽取将产生一个术语集合，本步骤的任务只考虑抽取出的术语关系在全部术语关系中的准确度。如果某一部分的术

语关系明显缺失，则对上一步文本相关度的设定以及术语抽取的原则加以调整。

（5）应用命中率分析。应用命中率分析的主要任务是分析出本体在实际应用时的工作状态。如果命中率不高，则分析是哪一部分的命中率缺陷，从而调整对应模块的权重。

九、本体的评价标准设计

通过上述对本体构建过程分析方法的设计，本文提出了本体评价的几个重要指标：

（1）术语抽取准确率；

（2）术语关系抽取准确率；

（3）应用命中率。

3.1.4 本体构建平台选择

本体在信息化、智能化领域具有很高的地位，本体的应用与构建也一直是研究热门。为了方便进行本体构建工作，世界上的研究机构开发了各种各样的领域本体构建工具[64]。常见比较成熟的本体构建工具有：OntoEdit，webOnto，protégé。表 3-1 对这三种不同的构建工具进行了对比：

表 3-1 不同本体构建平台的对比

评价指标	OntoEdit	protégé	webOnto
获取方式	免费	开源	只能在线免费使用
可视化视图	无	有	有
自带本体库	有	有	有
存储方式	文件	文件/数据库	数据库
操作难度	中等	简单	中等

由表 3-1 可知，protégé 能同时满足本地使用、支持开源、具备可视化功能并具有灵活的存储方式等要求，另外两种或多或少有所欠缺，因此本文在构建本体时选择 protégé 平台。

protégé 是由斯坦福大学医学院的医学情报研究组开发的，以 ODBC（Open Database Connectivity）为基础的，支持类、继承、模板和实例等知识表示要素的一个本体构建工具[65]。protégé 的特点如下：

（1）拥有可视化、图形化的用户界面，使用简便；

（2）支持 Unicode 字符集的输入，解决了跨语言、跨平台开展文本转换及处理的难题；

（3）可以免费下载、安装与使用系统软件及插件；

（4）支持 RDF，RDFS，OWL 等多种本体描述语言；

（5）系统支持本体在系统外进行编辑和修改，极其方便人工干预本体构建；

（6）protégé 开放源代码，并提供 API，便于用户在其他软件上集成使用。

protégé 开发界面如下所示：

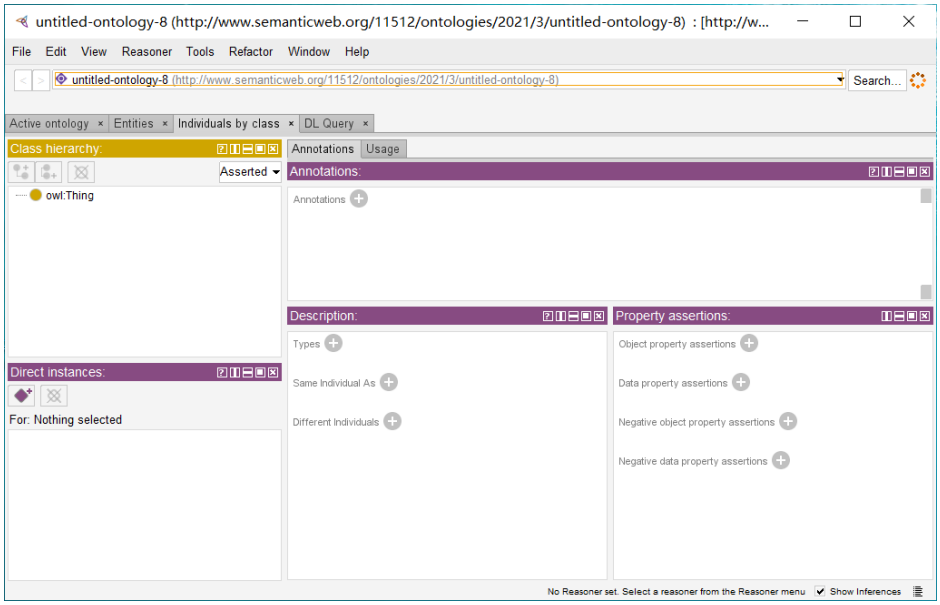

图 3.2 protégé 开发界面

3.2 工程测试技术本体构建过程

一、数据源选择

通过本文对于领域本体构建方法的研究，在进行本体构建之前，首先应选择所构建本体的范围，从而获得合适的数据源。针对工程测试技术而言，数据源的存储格式可大致分为以下三种类型[66]：

1、结构化数据源

结构化数据源一般是以关系数据库的形式存在的，这类数据源是经过人工逻辑处理的高质量的数据源，但是由于此类数据源的关系结构固定并且其中只能存

储一定量的信息，从而导致术语抽取和术语关系抽取过程中具有局限性并且过度依赖于数据源的质量。

2、半结构化数据源

半结构化数据和结构化数据类似，具有结构性，但是他的结构化程度低于结构化数据，具有可变化的数据结构。例如，XML 是一种典型的半结构化数据，XML 中具有不同的节点，而每个节点可以存储不同类别的数据信息。半结构化数据的动态特性更有利于术语的抽取和术语关系的确定，但是其查询率比较低。

3、非结构化数据源

非结构化数据是不具有固定格式的，例如，纯文本、视频、图片都是非结构化数据。非结构化数据在增长速度上远远高于结构化数据和半结构化数据，据调查显示，未来十年中所新增的数据之中，90%是非结构化数据，可以说非结构化数据是比较热门的数据类型。并且非结构化数据中包含更加隐晦的本体关系，基于非结构化数据将构建出更加全面客观的领域本体，质量更高。在非结构化数据中，纯文本是一种大量存在的类型，其中有着大量的信息并且方便获取，同时，大量的文本处理、自然语言处理技术可以帮助进行本体的构建与更新。

综上所述，本文将工程测试技术本体构建的数据源选择非结构化数据中的纯文本。为了减少本体构建对于人工的依赖程度，本文的语料库的获取是从维基百科抽取到的工程测试技术的领域网页语料库，抽取文档数为 5000 篇，以相同方式获取其他工科设计技术领域文档 600 篇，将之转换为 txt 文件格式；

二、数据处理

（1）文本分词

为了构建工程测试技术领域本体，需要对文本进行分词处理。本文仍然使用前文选择的中文分词工具—ICTCLAS 分词系统。在构建工程测试技术领域本体时，为提高本体识别的准确率，构建一个自定义领域词典来提高专业领域概念的分词的准确性。本文在明确本体构建的目标和范围的基础上，根据工程测试技术的相关知识和词汇的特征，通过人工获取的方式选取了 400 个工程测试技术常用词汇主要为工程测试技术的理论基础、方法策略、应用对象、应用工具等方面构建了工程测试技术领域词典。下表 3-2 为领域词典中的部分词汇。

表 3-2 测试技术领域词典

对象	理论	技术方法	工具
执行机构	虚拟理论	信号处理	速度传感器
机械手	信号描述	信号调理	激励装置
机器人	离散频谱	信号处理	转换装置
观察对象	傅里叶理论	显示记录	力传感器
传动装置	时域分析	传输	温度
…	…	…	…

导入以上的自定义词典后，使用 ICTCLAS 分词系统进行分词，分词算法流程如图 3.3 所示：

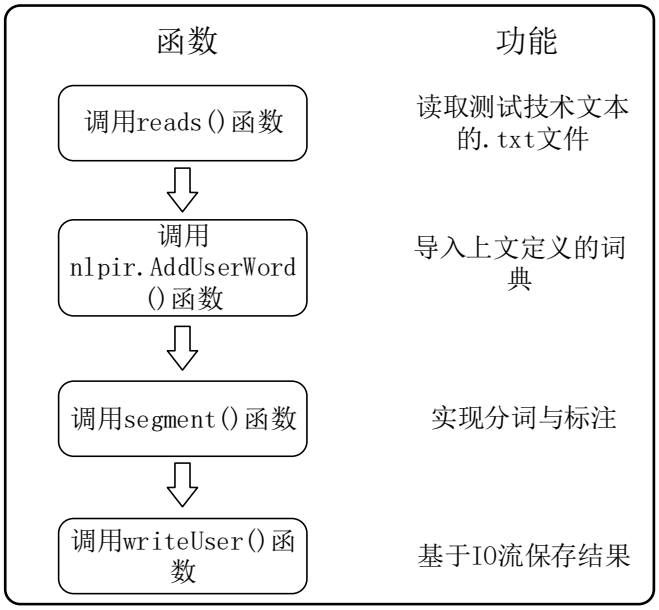

图 3.3 分词算法核心步骤

分词结果包含以下内容：

（1） 候选术语；

（2） 分隔符“/”；

（3） 词性。

具体分词结果如下图 3.4 所示：

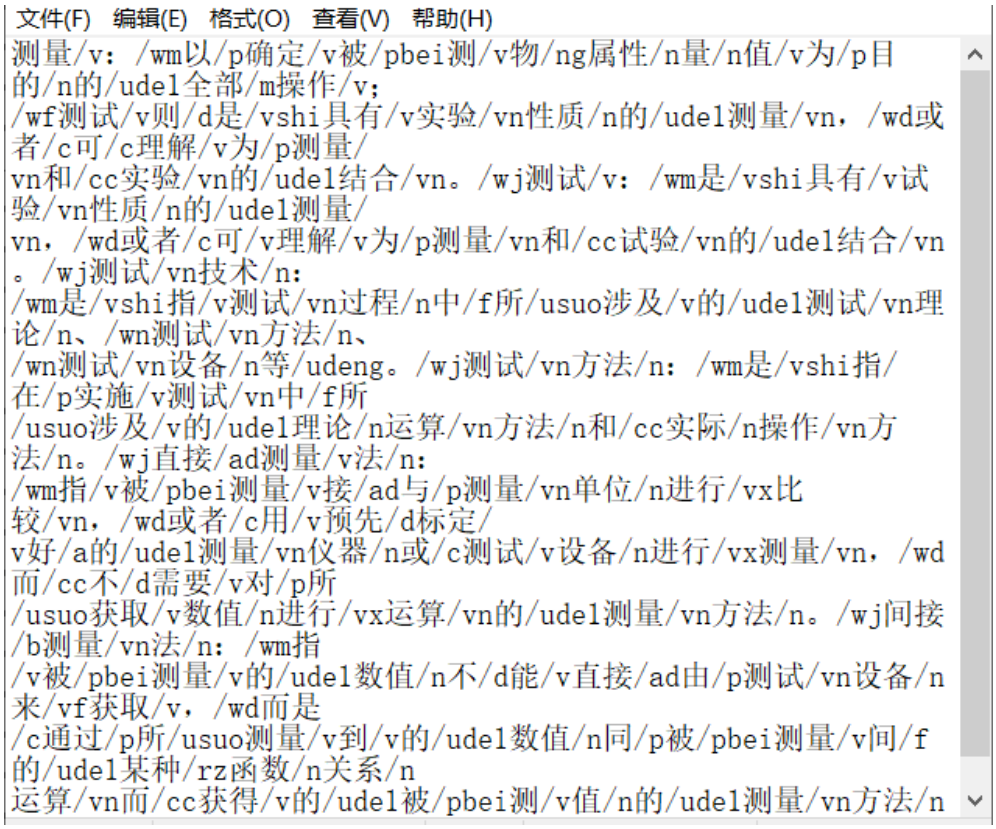

图 3.4 分词部分结果展示

（2）停用词过滤

经过分词处理后的工程测试技术文本中包含的大量词汇中有很多与主题无关，比如一些助词“的”、副词“一直”、“接着”，标点符号等，这些词就是工程测试技术文本中的停用词。本文将使用哈工大停用词表和百度提供的停用词表对处理后的文本进行过滤，将文本语料库中的停用词去除，以便于提高术语抽取及关系抽取的精度。

经过上述对于工程测试技术文本进行核心词典构建、分词、停用词过滤等一系列操作，得到工程测试技术本体候选术语集合，部分候选术语如下表3-3所示：

表 3-3 部分候选术语

部分候选术语集合			
测量	虚拟理论	信号处理	函数
机械手	数值	设备	激励装置
试验	离散频谱	信号处理	转换装置
观察对象	傅里叶理论	显示记录	力传感器
传动装置	时域分析	传输	标定
…	…	…	…

三、逻辑定义

本模块对工程测试技术领域的主要逻辑关系进行分析，主要内容包括：工程

测试技术术语类别统计、工程测试技术术语关系统计、工程测试技术应用试验分析、工程测试技术术语关系的确定以及形成逻辑集合等五个部分。

第一，候选术语类别统计。工程测试技术领域本体的主体部分便是术语集合，因此，术语的类别的统计是逻辑定义的第一步。首先，将经过数据收集步骤和数据处理步骤的语料文本作为输入，进入一个正则表达式统计术语类别的算法；其次，将每个候选术语的频率组成结果集；最后，将同一类别的术语组成同一个集合。这样，经过统计，工程测试技术文本库中候选术语的类别主要为：动词、名词+动词、名词+动词+名词几种结构。

第二，候选术语关系统计。首先，将经过术语类别统计的术语集合语料文本作为输入，进行术语关系类别的统计；其次，将每组术语之间的关系的频率组成结果集；最后，将同一类别的术语组成同一个集合。这样，最终获得工程测试技术文本语料库中术语关系的类别组成主要为：父子关系、顺序关系以及依赖关系。

第三，应用试验分析。经过上述两个步骤之后，在理论上工程测试技术领域本体构建所需要的的候选术语集合以及候选术语关系集合就可以确定下来。后续将根据应用的效果对本体质量进行评价。

第四，形成逻辑集合。在经过上述三个步骤之后，逻辑集合便可以确定。本步骤的主要任务是将术语之间的关系进行形式化的记录和持久化的保存。

四、术语抽取

工程测试技术领域本体构建的一个主体部分就是术语抽取。主要步骤如下：

（1）根据对比语料库，将领域语料库中与对比语料库中出现的高频重合词汇进行剔除，词频设定为 10；

（2）根据所得文本语料库，进行人工标记，得出 896 个术语，形成术语集合 A；

（3）使用本文的基于综合特征的术语抽取算法进行术语抽取，其中阈值 x 设定为 4.0，y 设定为 2.0，z 设置为 3.0。得出术语集合 B。

（4）将所得术语集合 B 与集合 A 进行比较，得出准确率，召回率和 F 值；

五、术语关系抽取

首先，确定剩余的合格术语，本文将上述处理过程中获得的术语集合去除只出现一次的术语集合之后的术语集合作为候选术语；其次，对文本集合进行遍历，

获得出现候选术语共现现象的句子；最后，本文提出术语关系候选词对的筛选原则：包含两个及以上候选术语的句子作为存在术语关系的句子，而这些对应的候选术语就组成了术语关系候选词对。本实验最终总共形成了 84526 个术语关系对。下一步就行从这些术语关系对中区分哪些术语关系对包含关系，以及包含什么样的关系。

1、规则分析

本文的规则分析主要完成对同义关系或者并列关系的分析提取。同义关系和并列关系符合的三种规则在第三章已经详细说明在此不再赘述。规则匹配这种方法具有高精确率的特点，本文对同义关系和并列关系的提取使用规则匹配的方式，这种方式为本体构建的高精度要求提供了基础。

2、关系过滤

本文按照公式 2.6 对术语关系对进行过滤。取 0.3 作为阈值，sim 小于 0.3 的时候往往不存在紧密关系。以这种方法来过滤掉相似度低于 0.3 的术语关系对。

3、依存解析

本文在构建工程测试技术本体时，对每一个句子进行依存解析，并通过提取实例的深层句法结构特征来为聚类做准备。

4、术语关系特征选取

本文的聚类方法按照第三章提出的方法步骤进行：首先，第一次聚类时将取三个特征：术语、术语词性、句子长度、依存路径长度；第二次则选取：依存路径长度和中间词特征。

5、K-means 聚类

基于以上的特征，最后使用 K-means 聚类进行术语关系的抽取，最终得出术语主要关系类型以及占比如表 3-4 所示：

表 3-4 测试技术本体主要术语关系

关系类型	占比
父子关系	0.25
顺序关系	0.16
依赖关系	0.18
同义关系	0.09
整体与部分关系	0.24
并列关系	0.08

六、合成本体

在对工程测试技术领域语料库进行术语抽取和术语关系抽取步骤完成之后，

本文将基于 protégé 平台进行工程测试技术的领域本体合成。主要步骤如下图所示：

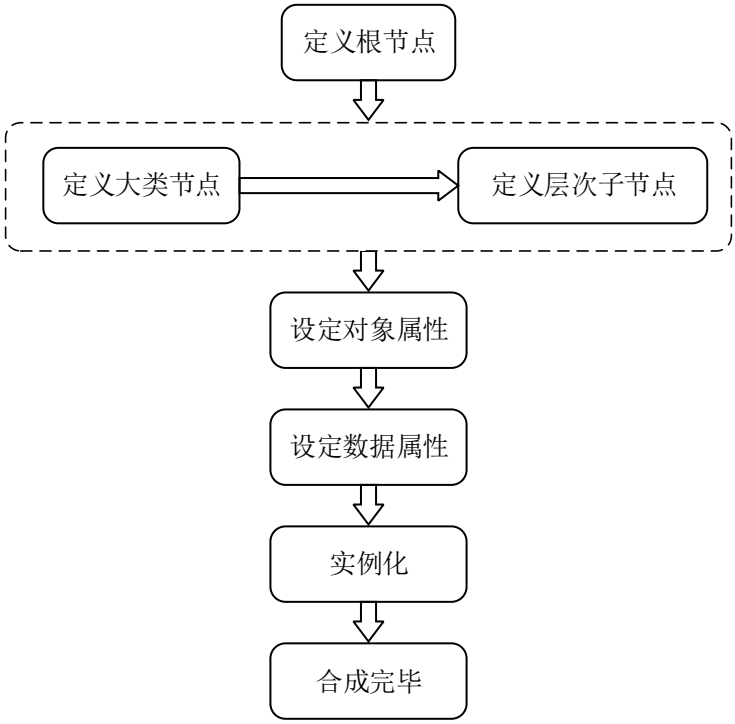

图 3.5 本体合成步骤

（1）根节点的定义。

本步骤将“工程测试技术”作为领域本体的根节点，其父节点是 protégé 提供的统一的父节点 thing 节点；

（2）核心大类节点以及层次节点的定义；例如，“激励装置”、“温度传感器”这些都属于此范畴；本文将工程测试技术本体的根节点的子类定义为大类，含义是，工程测试技术的核心类别，其再下一层的子节点全部属于核心大类，下文简称大类。大类包括：理论、测试方法、测试工具、测试经验、应用方式、测试目的以及测试意义。工程测试技术本体大类定义完成后，将术语抽取以及术语关系抽取的结果转化为实体和实体的层次关系，并使用 Entities>Classes 下的添加子类功能将术语间的层次关系建立完整。图 3.6 是工程测试技术本体大类和本体的层次结构的定义结果：

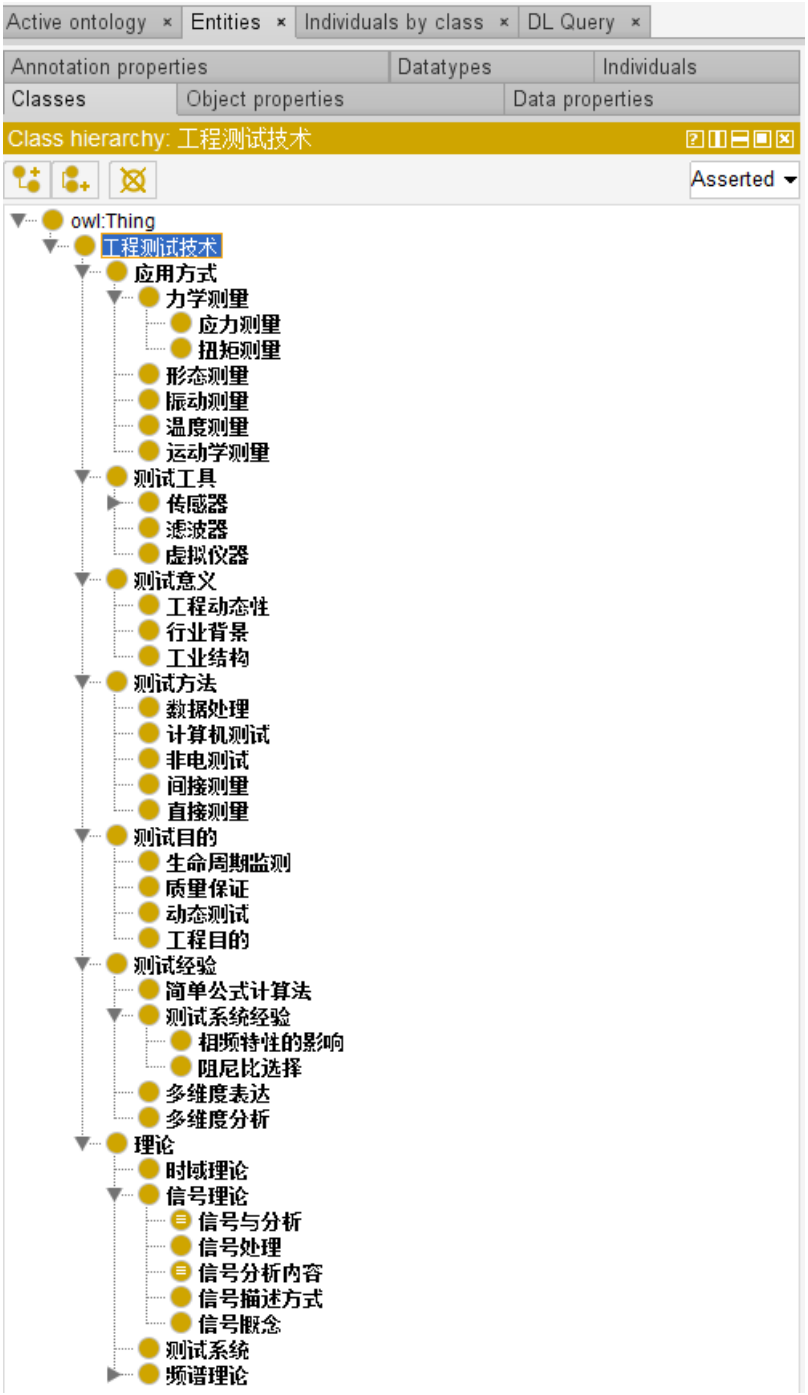

图 3.6 工程测试技术本体大类

（3）设置对象属性

经过以上步骤，本文已经得到了工程测试技术领域本体的术语以及术语之间的层次关系，接下来，需要对不同的术语实体之间的非层次关系进行输入，在 Protégé 中非层次关系进行输入以设置对象属性的方式体现。需要注意的是，非层次关系的输入关系层次只能定义在非叶子节点的级别上，后续叶子节点会对该对象属性进行直接使用。非层次关系主要包括：应用关系、依赖关系、辅助关系、

同义关系等。Protégé 提供了规定的实体关系，用户也可以通过设置对象属性的方式来自定义实体关系，其中同义关系即为 Protégé 规定的，应用关系术语用户自定义的。

对于 Protégé 规定的实体关系，设定方式如下图所示：

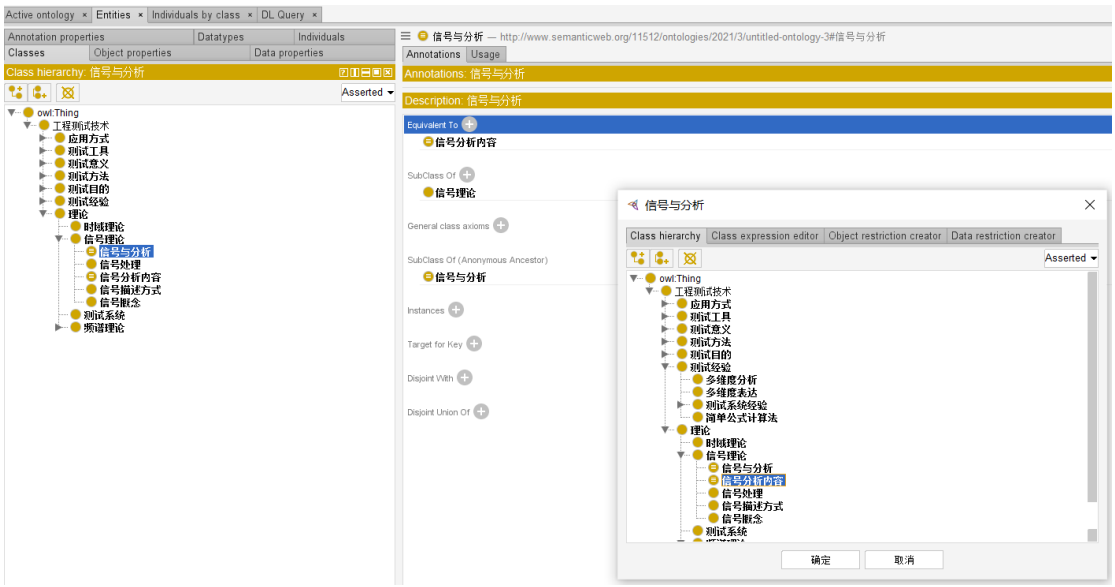

图 3.7 protégé 提供的实体关系

首先将操作空间选中在信号与分析这个实体类，将后续操作与该类绑定。然后，选择 Description 空间，对应位置的扩展符号右侧将提示当前本体的实体层次关系，根据提示选择对应的同义关系类，确定选择后即设置成功。

对于用户自定义的实体关系的创建而言，设定方式如下图所示：

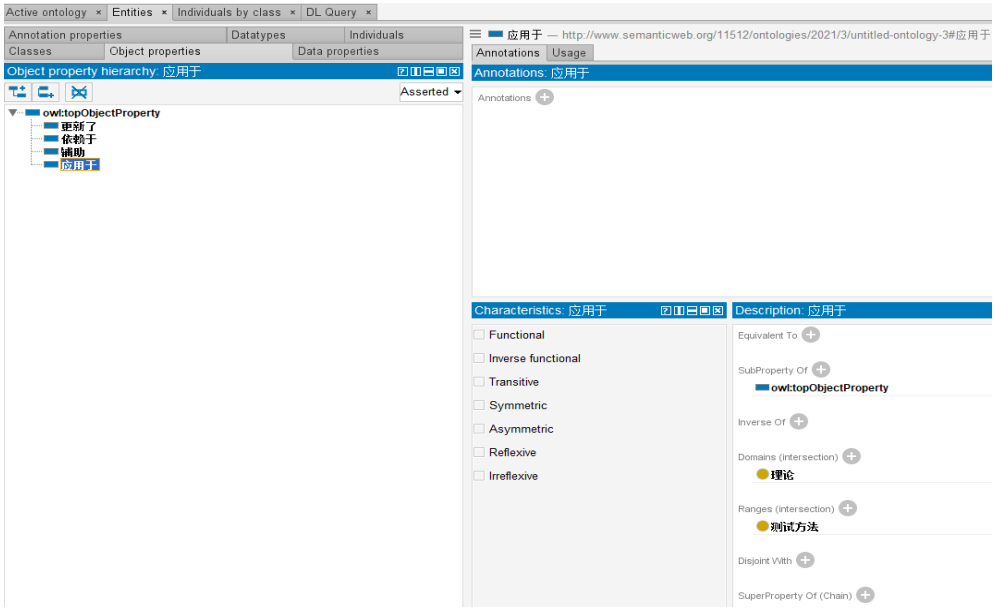

图 3.8 自定义关系设定方法

首先，在 Object properties 操作空间下，添加 sub-properties；然后，在 Description 空间下设定 Domain，即定义该实体关系的主动方，同时设定 Ranges，即定义该实体关系的被动方；最后，在下一步实例化中，创建叶子节点后，选择对应的自定义实体关系，完成工程测试技术本体的关系网络的创建。

（4）设定数据属性；以传感器节点为例，“别名”、“型号”、“原理”以及“使用方法”等，即节点的属性。设定方式如图 3.9 所示：

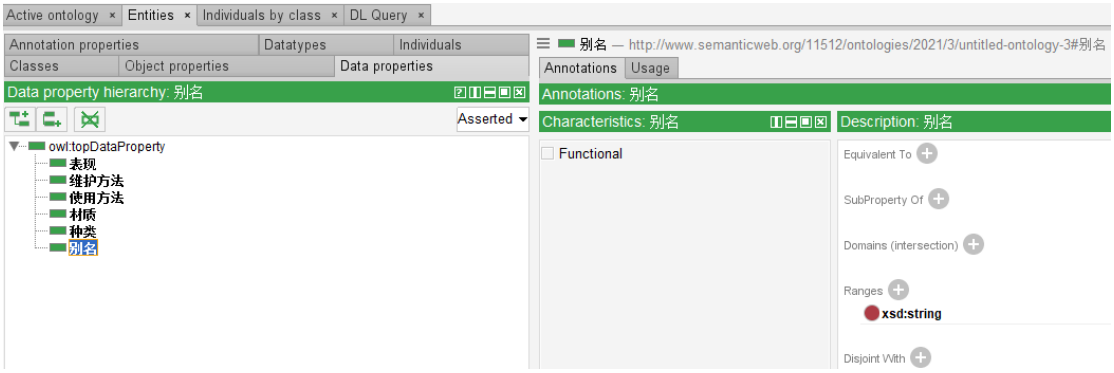

图 3.9 数据属性设定方式

（5）实例化；例如，具体的“xx 型温度传感器”就属于实例或者称之为叶子节点。此类节点需要注意的是，要在上述设定对象属性和数据属性的基础上，选择对应的属性来与其他实例产生关系映射。设定方式如图 3.10：

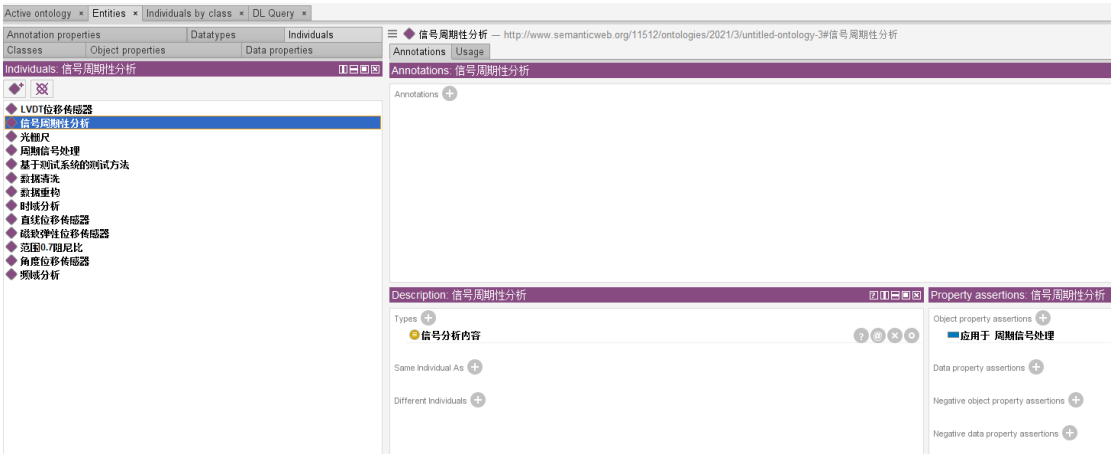

图 3.10 本体实例化方法

首先进入 individuals 空间，新建本体实例；然后，在类别关系空间根据提示选择正确的层次关系类别；最后，选择需要的对象属性以及对象属性的被动方，即完成本体实例化的步骤。

经过上述对于工程测试技术领域本体的合成，在 Protégé 中以 OntoGraf 可视化功能进行工程测试技术领域本体的展示，如图 3.11 所示：

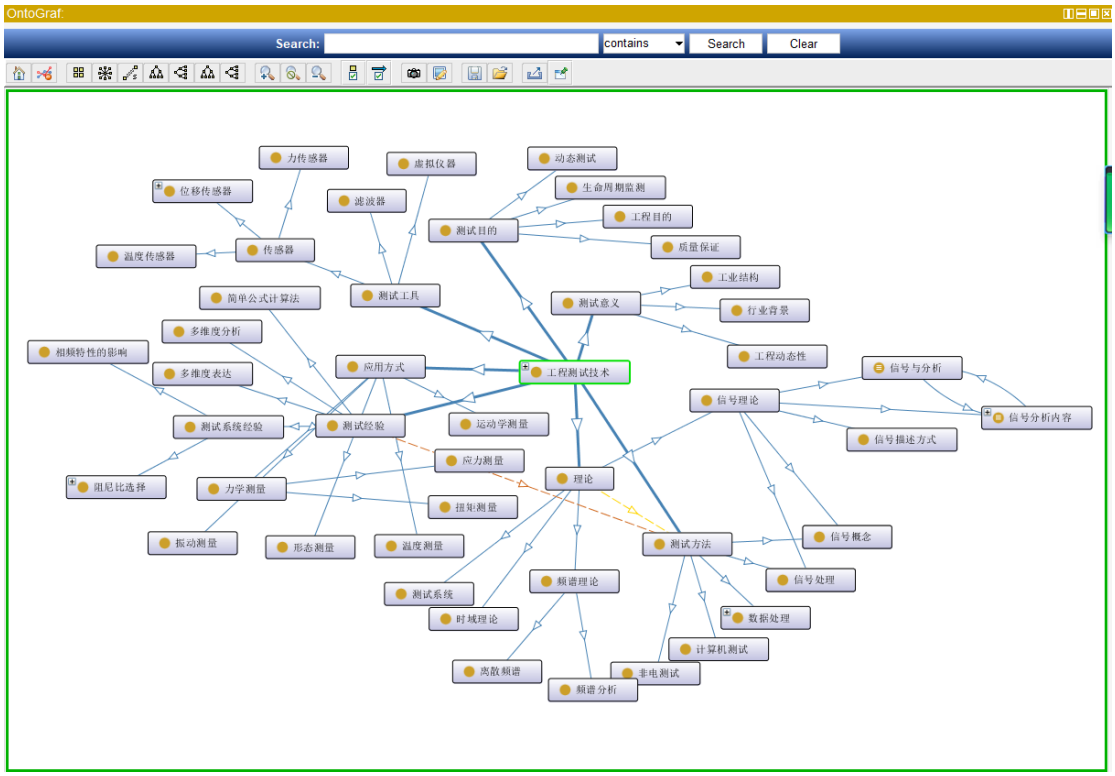

图 3.11 工程测试技术领域本体

上图为工程测试技术领域本体主要层次结构，是一个整体的网络结构，针对七个主要大类，本文给出了 “直线式位移传感器”和 “信号分析理论”两个实例的本体结构进行展示：

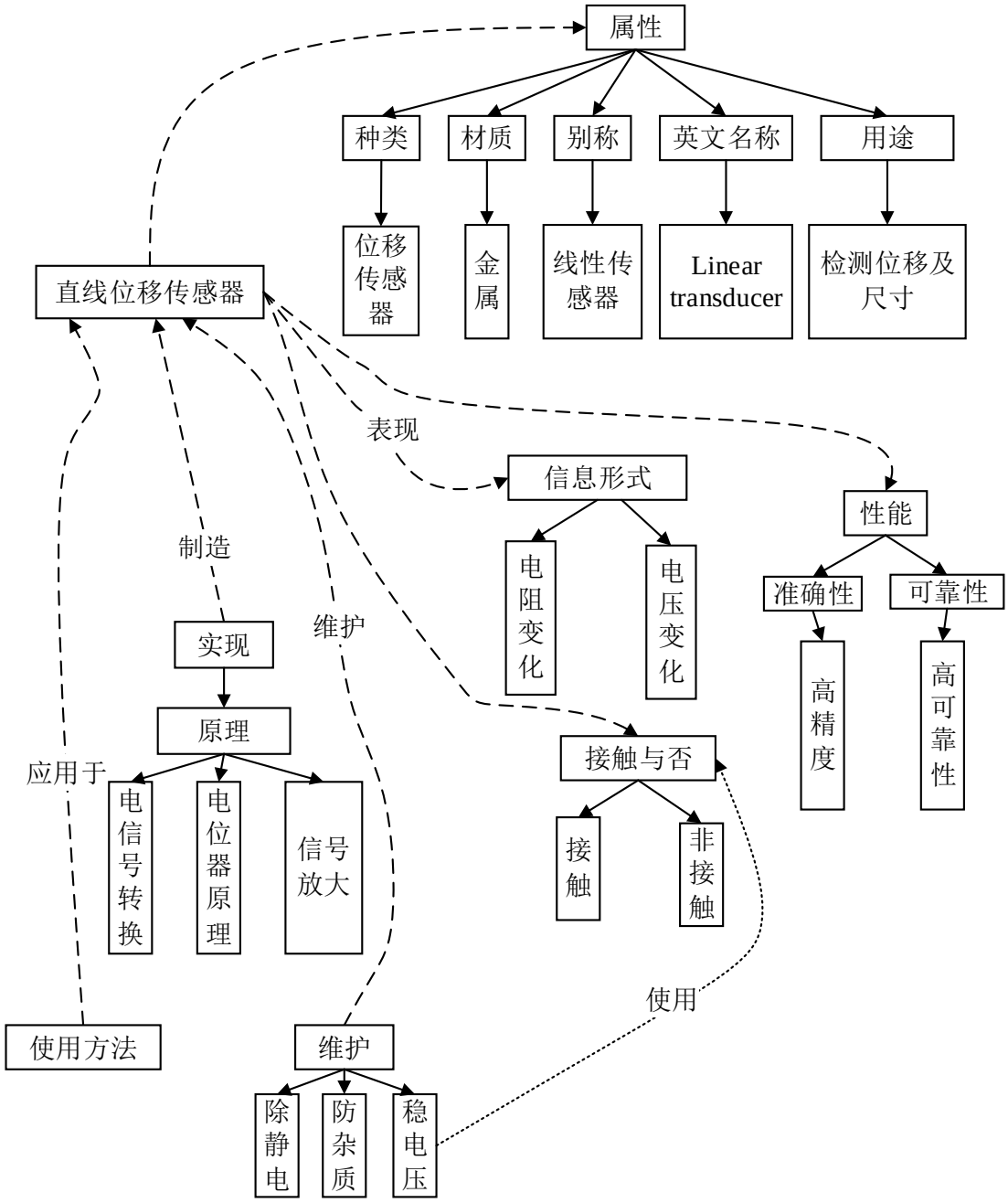

图 3.12 直线位移传感器本体

图 3.12 是直线位移传感器本体，对象属性包括：应用于、制造、维护、表现；与该类别（位移传感器）有非层次关系的类有：使用方法、原理、维护、信息形式、信息形式，这些类的内部又包含各自的对象属性、数据属性以及其他层次关系，在此不做展示；数据属性包括：性能和属性。性能下一层分为：准确性和可靠性两个实体类，属性类包含种类、材质、英文名、别称及用途五个数据属性。

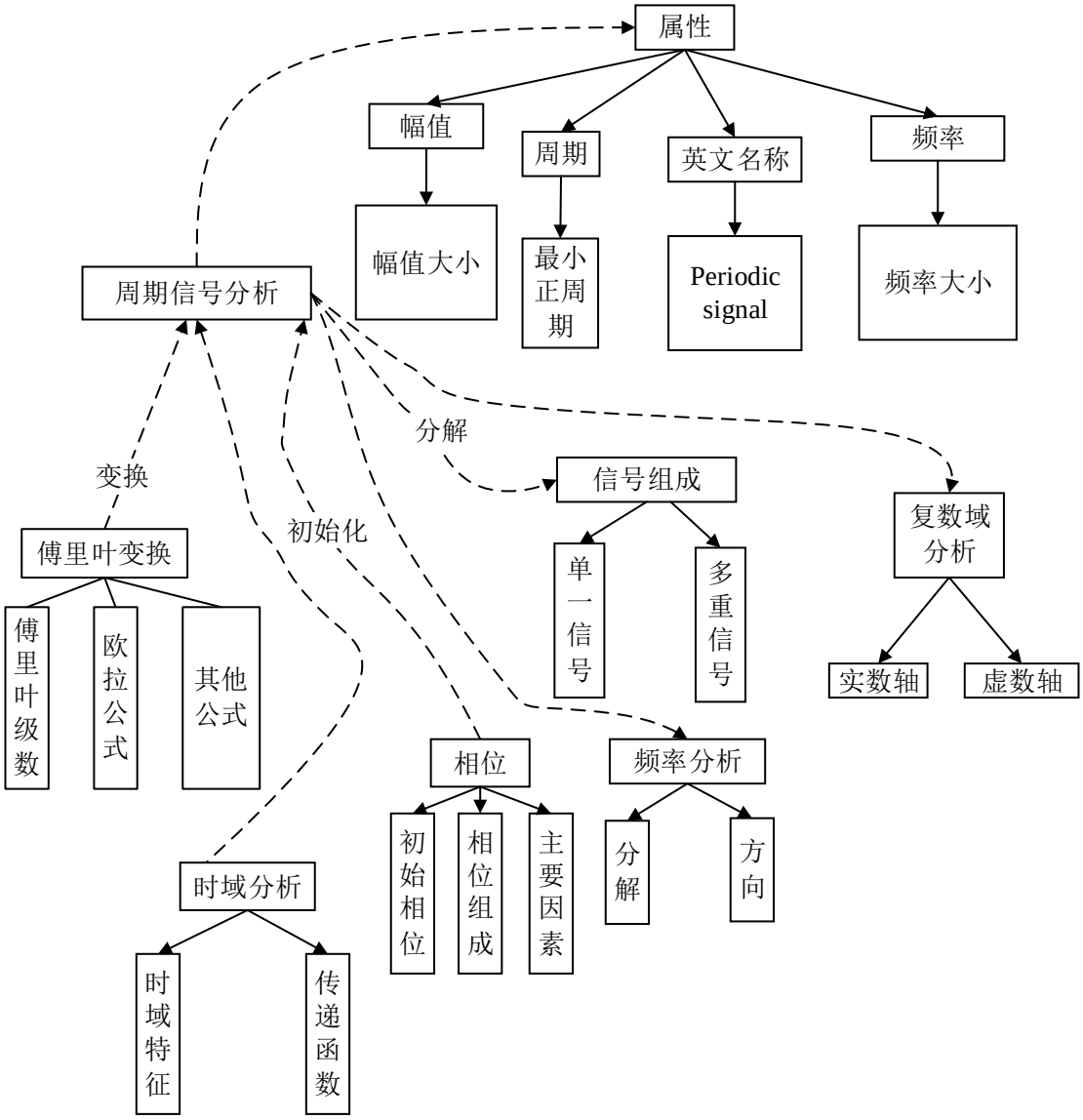

图 3.13 周期信号分析理论本体

图 3.13 是周期信号分析实例本体，对象属性包括：变换、初始化、分解、属性。与其具有关系的类别包括：傅里叶变换、时域分析、相位、信号组成、频率分析、属性以及复数域分析。

七、形式化编码

在学者们对本体相关的理论与技术的研究进程中，本体描述语言也在发生着不断地变化。最初的 HTML 语言是描述语言的典型案例，但是其无法描述丰富的语法和数据语义，因此，学者们使用 XML 语言代替 HTML 语言，以满足开发者进行特殊领域的对数据的定义以及扩展，并为特定的树状结构提供序列化语法。但是 XML 语言也存在自身的缺点：（1）互操作性不足；（2）无法实现元数据建模，故 W3C 组织为解决此问题开发出 Web 元数据所特有的规范—资源描述框架

（Resource Description Framework, RDF），随后又提出了新的规范--资源描述架构规范(RDF Schema, RDFS)，这个规范是用以扩充 RDF 数据的。随着语义网的发展，Web 本体语言（Web Ontology Language,OWL）也随之诞生。与 XML,RDF,RDFS 相比，OWL 能清晰的表达术语和术语之间的关系，尤其是对术语的含义和复杂的术语关系的表达更能凸显其优势，即具有更多的机制来表达复杂的语义网络。因此在语义 web 中，更具有广泛用途的就是 OWL 语言。本体的层次化关系如图所示：

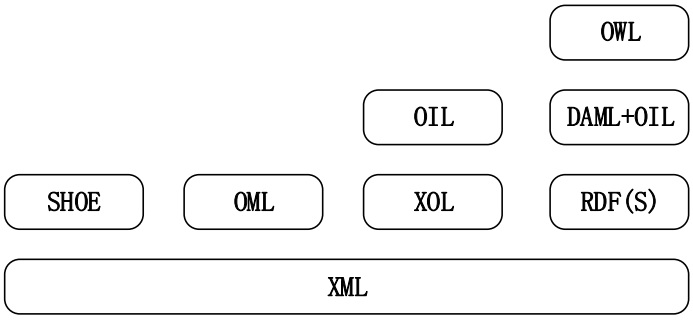

图 3.14 本体描述语言的层次化关系

本文对于工程测试技术本体采用 OWL 形式化语言进行描述，部分代码如下所示：

```
<?xml version="1.0"?>
<rdf:RDF      xmlns="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#"
      xml:base="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3"
      xmlns:owl="http://www.w3.org/2002/07/owl#"
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:xml="http://www.w3.org/XML/1998/namespace"
      xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
    <owl:Ontology rdf:about="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3"/><!--  http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#传感器 -->
    <owl:Class rdf:about="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#传感器">
<rdfs:subClassOf rdf:resource="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#测试工具"/></owl:Class>
    <!--  http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#位移传感器 -->
    <owl:Class
```

rdf:about="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#位移传感器">
<rdfs:subClassOf rdf:resource="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#传感器"/></owl:Class>
<!-- http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#信号分析内容 -->
<owl:Class rdf:about="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#信号分析内容">
<rdfs:subClassOf rdf:resource="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#信号理论"/></owl:Class>
<!-- http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#信号描述方式 -->
<owl:Class rdf:about="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#信号描述方式">
<rdfs:subClassOf rdf:resource="http://www.semanticweb.org/11512/ontologies/2021/3/untitled-ontology-3#信号理论"/></owl:Class>

3.3 本体持久化

在完成本体的构建之后，本文对工程测试技术领域本体的持久化方式进行了研究。工程测试技术领域本体持久化的主要内容是将该本体持久化到磁盘上，本文主要采用基于数据库的方式和基于文件管理的方式。

3.3.1 持久化方式对比

一、文件管理方式

若将本体以文件的形式存储于计算机硬盘上，用以本体的更新以及复用本体。此时，数据以文件的形式组织与保存。在需要使用该文件时，开发者可以调用文件系统提供的操作命令来建立和访问该文件。此类型的数据持久化方式具有以下特点：数据可以长期保持、文件系统管理数据、数据共享性差，冗余度大、数据独立性差[67-69]。

1、数据可以长期保存

计算机大量数据处理数据，数据需要长时间保留在外存上反复进行查询、修

改、插入和删除等操作。

2、由文件系统管理数据

由专门的软件即文件系统进行数据管理，文件系统把数据组织成相互独立的数据文件，利用“按文件名访问，按记录进行存取”的管理技术，可以对文件进行修改、插入和删除的操作。

3、数据共享性差，冗余度大

在文件系统中，一个（或一组）文件基本上对应一个应用程序，即文件仍然是面向应用的。不同的应用程序具有部分相同的数据时，也必须建立各自的文件，而不能共享相同的数据，因此数据的冗余度（redundancy）大，浪费存储空间，而且由于重复存储、各自管理，容易造成数据不一致，增加了数据修改和维护的难度。

4、数据独立性差

文件系统中的文件为某一特定应用服务，文件的逻辑结构对该应用程序来说是优化的，所以要想对现有的数据再增加新的应用是很困难的，系统不易扩充。

二、数据库方式

若将工程测试技术领域本体中的每个节点作为一张表进行存储，每个表格中的字段是由该节点的属性以及与关系其他节点的关系组成，由外键建立起与其他表格之间的联系，此种方式可用于对本体的快速检索以及应用。相比于文件管理的方式，数据库持久化本体数据具有明显的优点，其主要特点如下：

1、数据结构化

数据库系统实现整体数据的结构化，也是数据库系统与文件系统的本质区别。不仅数据内部是结构化的，而且整体也是结构化的，数据之间是有联系的，而文件系统只是内部有结构，但整体无结构，记录之间没有联系[70]。

2、数据的共享性高，冗余度低，易扩充

针对工程测试技术本体的数据库从整体角度看待和描述数据，数据不再面向某个应用而是面向整个系统，因此数据可以被多个用户、多个应用共享使用。数据共享可以大大减少数据冗余，节约存储空间，还能避免数据间的不相容性和不一致性。

3、数据由 DBMS 统一管理和控制

数据库的共享是并发的共享，即多个用户可以同时存取数据库中的数据，甚至可以同时存取数据库中同一个数据。DBMS 可提供领域本体数据的安全性保护、完整性检查、并发控制、数据可恢复等功能。

3.3.2 本体持久化方式设计

综上所述，本文将采用文件管理与数据库管理工程测试技术本体两种方式相结合的方法。文件管理主要侧重对于本体的复用以及更新时的操作，数据库方式将更加侧重于在基于本体实现的应用时，提升领域本体的获取效率。这样二者结合，就可满足本文对于领域本体构建以及应用的需求。

3.4 本章小结

本章基于上文提出的本体构建方法以及本体自动构建算法，构建出了工程测试技术本体。在构建本体的过程中，首先，进行了工程测试技术本体构建原则的分析；其次，进行术语抽取和术语关系抽取；再次，对抽取出的术语和术语对关系进行了本体的合成。最后，借助 Protégé 工具对工程测试技术领域本体来进行可视化展示，使用文件管理以及数据库管理相结合的方式进行工程测试技术本体的持久化。本章构建的工程测试技术本体为后续实现信息检索系统提供了底层的逻辑支持。

第四章 语义检索模型设计

对于一个检索模型而言，其主要目的就是为某个领域或者多个领域的检索系统提供统一的范式或者逻辑，并且不同的检索模型有着不同的侧重点：针对海量数据、针对特定数据类型的检索以及针对检索的智能性等等。

本章提出的基于本体的需求融合信息检索模型，将引入用户需求模型，在对于用户需求的模型构建以及更新的基础上，将之融合在检索过程中以达到用户的个性化检索需求。

4.1 用户需求模型

在工科相关领域的工程作业中，数据的种类繁多，并且由于工程是一个多学科交叉的对象，在构建领域本体时，不可能将所有的术语之间的关系全部整合在一起并持久化。不同的用户对于相关信息的检索将存在不同的个性化的需求，这为检索实现语义功能带来了阻碍[71-73]。为了解决这个问题，本文将提出一种用户需求模型，并在用户需求模型中加入工程测试技术领域常见的关系：因果关系，对立关系，互补关系，互利关系。本文将用户检索模型与经典信息检索模型相结合，提出一种改进的信息检索模型，结构如图 4.1 所示。

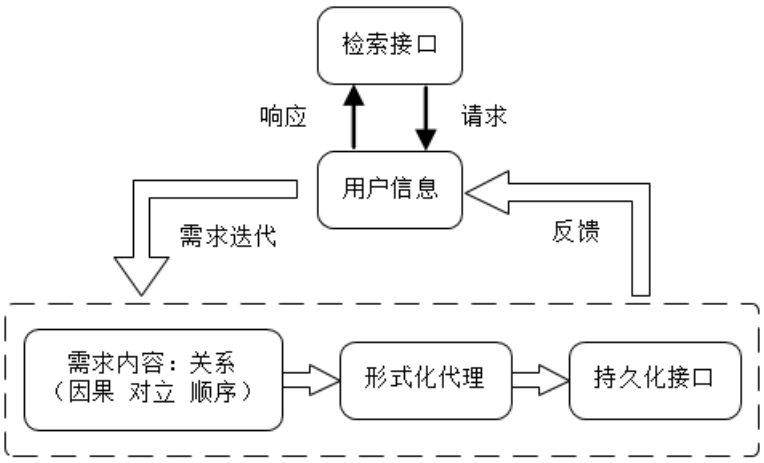

图 4.1 用户需求模型

4.2 用户需求模型的内容

对于使用一个检索系统的不同用户而言，从数据集合中返回的相关数据集合并非一定要面面俱到，他们只要求接收的数据集合中尽可能多的是自己所关心的内容。那么，如何将用户需求的内容以特定的方式和逻辑保留下来是一个重要的问题。

用户需求的内容包括：需求内容（关系）、形式化代理、持久化接口、用户信息以及检索接口。

一、需求内容

目前的个性化检索模型中，对于用户检索的个性化的依据主要是用户的检索记录。但是检索记录是很笼统的一个对象，要基于用户检索记录来达到个性化检索，首要任务就是对用户行为的关键部分加以抽象。

本文提出的用户需求模型从用户行为中抽取出查询对象与相应结果数据集合之间的需求关系。确定这些关系的主要方式是通过网络资料的查阅以及通过典型工程领域人员的分析指导总结出的关系规则。这些关系包含以下三个部分：因果关系、顺序关系和对立关系。这三种关系是本文在分析了检索过程中查询对象与相关结果数据集合之间的关系之后总结出的，并且不同的用户对于检索的需求的差别主要是这三种关系。

二、形式化代理

用户需求模型中的形式化代理模块的主要任务是将需求内容中的关系整合成为一个完整的、合理的并且可存储的结构[74]。本文提出的用户需求模型将需求内容和用户信息加以整合，以一个三元组$\{F, R, S\}$的形式来描述用户需求，其中 F 指代用户检索时输入的特征项，R 表示关系，S 表示系统反馈给用户的相关特征项。用户需求模型中的形式化代理模块的主要特点就是只对用户信息中关键点进行提取，而不关心和检索明显无关的其他因素。这些关键点主要包括：用户检索行为中的常用检索关系、用户检索行为中的检索关系的权重。

基于需求内容以及迭代的用户信息，本模型最终以本体的形式来对用户需求进行整合，并形成用户需求本体。

三、持久化接口

用户需求模型中的持久化接口的主要任务是将领域本体、用户需求本体以及

相关数据集合进行存储。在持久化到磁盘上的过程中应当注意：对于领域本体和用户需求本体应当开辟不同的位置加以存储。对于用户需求本体而言，不同的用户应当具备独立的用户需求本体，这里的用户需求本体只体现了不同用户的特性，而没有将所有用户的共有需求形式化，因此，不同用户的兴趣本体将占很少的内存。另外要注意的一个点是，在对用户需求本体进行持久化的同时，应当将本体中关系的权重同时更新，本模型将使用队列这种数据结构，将用户需求本体中的关系的权重加以持久化，以便给检索时关系权重的赋值提供依据。

四、用户信息

用户信息模块主要与两个模块进行对接：检索接口和经过形式化的需求。在一次用户的检索行为中，当用户信息模块第一次接收到来自检索接口的请求时，将根据用户的身份从现有的用户需求本体库中获取并返回响应的需求集合。第二次接收到检索接口传来的请求时（这时用户已经检索到相关的数据集合），用户信息模块根据用户此次的查询特征项以及返回数据集合的特征项之间的关系，更新用户需求本体中的关系的权值。之后，将使用持久化模块对此次更新进行同步。这样，随着用户检索次数的增加，检索的语义性将越来越强并更加能满足用户的需求。

五、检索接口

检索接口的一部分功能是与上述的用户信息模块进行交互，在此不再赘述。检索的另一部分功能是为作为信息检索模型的一部分而承担的任务。在检索模型中，用户在界面输入查询指令后，将在本体的语义扩展的帮助下，返回相关的数据条目，此时检索接口将从用户信息中获取的用户需求发送给检索模型，检索模型将依据对应的用户需求对检索结果集加以修正，以达到用户需求融合于信息检索模型中的目的。

4.3 基于本体的需求融合语义检索模型

本文提出的改进的信息检索模型的目的包括两个方面：提高信息检索的语义性和增强用户自主性。基于以上两点提出对于信息检索模型的要求：

A：通过相关理论以及工具的整合，将数据集合的语义性体现出来；

B：能够使用户需求参与到信息检索的过程中；

C：模型的各模块之间有机联系，不重复，不缺失。

基于用户需求模型的建立，本文将依赖本体，并在用户需求模型的融合下，提出基于本体的用户需求融合语义检索模型，如图 4.2 所示：

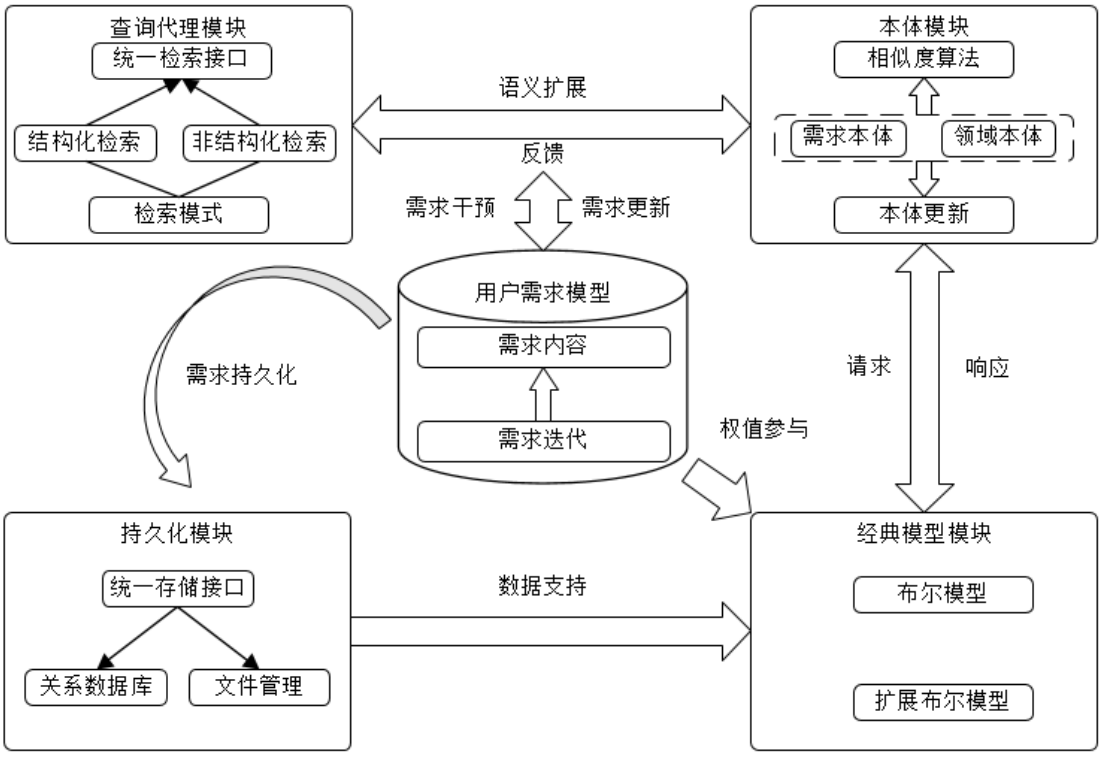

图 4.2 基于本体的需求融合语义检索模型

4.3.1 查询代理模块

查询代理模块是本文提出的基于本体的需求融合信息检索模型的第一模块，该模块通过 UI 界面与用户进行交互。对于用户而言，用户通过界面输入查询指令，这些指令通过查询代理模块的进一步处理之后进入到检索模型中。对于查询代理模块而言，它扮演着信息检索模型的接口的作用，将检索模型获取到的相关数据集合反馈到界面中。

基于本体的需求融合信息检索模型依据检索模式的不同，对用户的检索行为进行了不同的处理。它将检索模式分为两种：结构化检索和非结构化检索。对于这两种检索模式，将基于不同的经典信息检索模型进行检索。

对于结构化检索，其检索方式的输入必然是结构化的，即当用户的检索行为是输入结构化的检索指令时，基于本体的需求融合信息检索模型就判定其是结构

化检索。判断是否是结构化检索的关键一点是判断用户的检索语句中是否是以空格断开的，若用户的检索语句以空格隔开了多个关键词，那么本模型就判定其为结构化检索。对于结构化检索，本模型采用布尔模型进行检索。检索过程如下：首先，通过对检索语句的特征提取，得到了相关的特征项集合；其次，将之转换为析取范式；最后，根据析取范式和领域本体的语义相似度结果，检索出正确的数据集合。

对于非结构化检索，其检索方式当然是非结构化的，即当用户的检索行为是输入非结构化的指令时，基于本体的需求融合信息检索模型就判定其是非结构化检索。判断其是否是非结构化检索的依据刚好与结构化检索的判断依据相反：当用户的检索语句是以自然语言的方式来表达时，模型就判定其为非结构化检索。对于非结构化检索，基于本体的需求融合信息检索模型将采用需求模型与扩展布尔模型相结合的方式来进行相关数据集合的抽取。

4.3.2 本体模块

本体模块是基于本体的需求融合信息检索模型中实现检索的语义性的担当。它是相关领域以及用户需求的形式化表达，是查询代理模块进行查询扩展的重要依据。在本体模块的应用之前，需要对领域本体进行构建。在构建了领域本体之后，采用OWL语言进行本体的形式化的描述以及存储，将之持久化到数据库中。本体模块的主要应用逻辑就是：对查询模块的语义支撑，根据本体中的概念相似度计算为查询特征项的扩展提供依据。

领域本体的是基于本体的需求融合信息检索模型的语义性的主要承担者，它是相关领域中的类及关系等对象的形式化表达。在查询代理模块收到用户的查询指令后，为了实现语义检索，必须要进行查询扩展，而查询扩展的手段就是通过领域本体所包含的关系进行语义相似度的计算。所以，领域本体是查询代理模块的有效支撑。

用户需求本体的将参与到信息检索中的查询代理模块中，它将两次接受从查询代理模块传来的数据。当用户输入查询指令后，查询代理模块基于领域本体和用户需求本体进行基于语义的查询扩展，此时将扩展的特征项条目返回到UI界面，在用户进行简单的选择淘汰后，确定具体的特征词集合，依据该特征词集合，

得到信息检索的结果。此时，基于本体的需求融合信息检索模型将初始检索特征项与扩展后的特征项集合进行比较，得出此次检索应用到的关系集合，关系的种类从用户需求本体中获得。之后，将本次检索的相关的关系在数据库中存储关系权值的队列进行更新，每次给权值加 1 并重新排序后存储到队列中。在下次检索中，将根据需求关系的权值对检索过程的权值进行改进。

语义相似度的计算是查询扩展的重要依据。根据已经构建好的领域本体和用户需求本体，基于本体的需求融合信息检索模型将采用基于距离的语义相似度计算方法。由于领域本体和用户需求本体是以 OWL 语言进行形式化的表达，因此本模型采用 Jena 推理机对两个本体进行解析，对 Jena 推理机的介绍将在第五章进行。在此基础上，计算与查询特征项语义相似度达到阈值的相关概念。

4.3.3 持久化模块

对于持久化模块而言，其主要任务就是完成对领域本体、用户需求本体以及需求关系权值的持久化，图 4.3 展示了基于本体的需求融合信息检索模型的数据持久化方法。

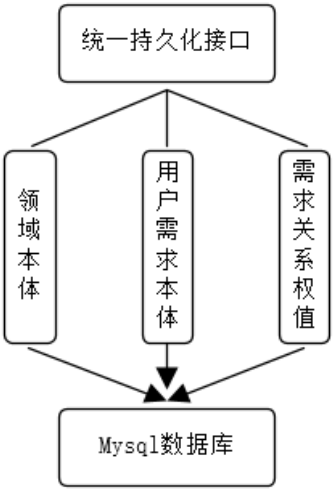

图 4.3 数据持久化方法

对于领域本体和用户需求本体而言，其持久化的方式都是将对应的 OWL 形式化表达的文件直接持久化到 mysql 数据库中。而对于需求关系权值而言，基于本体的需求融合信息检索模型将通过映射结构将其存储，每个关系类型将具备对应的权值。在每次的用户检索行为之后将对该存储部分进行更新，并将重新对关

系进行排序存储。

4.3.4 经典模型模块

经典模型模块是基于本体的需求融合信息检索模型的基础。对于结构化检索将应用布尔模型，对于非结构化检索，将基于概率模型的迭代思想，采用用户需求模型与扩展布尔模型相结合的方式进行信息检索。

对于结构化检索过程，基于本体的需求融合信息检索模型先将用户检索指令转换为标准析取范式，再根据析取范式进行正确结果集的提取。主要逻辑如下：

设文本集合 D 中某一文本 i，该文本可表示为：$D_i = (t_1, t_2, \dots t_m)$，其中，$t_1, t_2, \dots t_m$为标引词，用以反映 i 的内容，用户检索式如下：$Q_j = (t_1 \wedge t_2) \vee (t_3 \wedge (-t_4))$。对于该检索式，系统响应并输出的一组文本应为：它们都包含t_1和t_2，或者包含t_3但是不包含t_4。

对于非结构化检索过程，基于本体的需求融合信息检索模型将应用扩展布尔模型进行信息的检索。扩展布尔模型的检索式分别为：

$$Q_{\vee(p)} = (t_1, a_1) \vee (t_2, a_2) \dots \vee (t_n, a_n) \tag{4.1}$$

$$Q_{\wedge(p)} = (t_1, a_1) \wedge (t_2, a_2) \dots \wedge (t_n, a_n) \tag{4.2}$$

扩展布尔模型的文本和查询的相似度定义为：

$$sim\big(d, Q_{\vee(p)}\big) = \left[\frac{a_1^p d_1^p + a_2^p d_2^p + \cdots a_n^p d_n^p}{a_1^p + a_2^p + \cdots + a_n^p}\right]^{\frac{1}{p}} \tag{4.3}$$

$$sim\big(d, Q_{\wedge(p)}\big) = 1 - \left[\frac{a_1^p (1-d_1)^p + a_2^p (1-d_2)^p + \cdots + a_n^p (1-d_n)^p}{a_1^p + a_2^p + \cdots + a_n^p}\right] \tag{4.4}$$

上式中，sim 表示相似度，d 表示特征项集合，$Q_{\vee(p)}$表示检索语句最终表达成“或”的形式，p 表示查询维度，取值范围为$[0, \infty]$。n 表示文本维度。a_1，a_2，…，a_n表示不同特征项的权值，取值范围为$[0, 1]$。$Q_{\wedge(p)}$表示检索语句最终表达成“与”的形式。

4.3.5 用户需求的融合

基于本体的需求融合信息检索模型中，用户需求将通过融合的方式对传统信息检索模型进行改进，主要体现在两个方面：一方面，用户在进行信息检索时将加入用户对于扩展词的简单的选择淘汰步骤，能够更好的增加语义检索的个性化。另一方面，在非结构化检索的模型相似度计算方面，本模型基于概率检索模型的迭代思想，将检索模型的权值确定的方式加以改进。采用上述提出的需求关系权值作为信息检索过程中特征项的权值，参与到相似度的运算过程，基于本体的需求融合信息检索模型给出以下公式计算相似度：

$$sim\left(d, Q_{\vee(p)}\right) = \left[\frac{c_1^p d_1^p + c_2^p d_2^p + \cdots c_n^p d_n^p}{c_1^p + c_2^p + \cdots + c_n^p}\right]^{\frac{1}{p}} \quad (4.5)$$

$$sim\left(d, Q_{\wedge(p)}\right) = 1 - \left[\frac{c_1^p (1-d_1)^p + c_2^p (1-d_2)^p + \cdots + c_n^p (1-d_n)^p}{a_1^p + a_2^p + \cdots + a_n^p}\right] \quad (4.6)$$

其中，c_1，c_2，…，c_n表示检索词与初始查询词之间的关系的关系权值，取值范围为$[0，1]$。该式将特征项与文本之间的相似度以 n 维向量的之间的距离来加以描述。此时，查询特征项是经过查询代理模块扩展的来的特征项集合，此集合中必然包含经过本体的语义相似度模块计算而新增的特征项，这些新增特征项与初始特征项共同构成了用户一次检索行为的特征项集合。此时，基于本体的需求融合信息检索模型将再次调用本体模块，查找新增特征项与初始特征项之间的关系。在此基础上，以持久化模块为数据支撑，获取对应的关系目前的权值，作为每个新增特征项的权值。

4.4 本章小结

本章提出了基于本体的需求融合语义检索模型。首先，设计了用户需求模型的整体架构以及主要模块，并对用户需求模型的内容进行了详尽的阐述；其次，将用户需求模型与本体融入了传统的信息检索模型内，提出了基于本体的需求融合语义检索模型，该模型主要包括以下五个模块：查询代理模块、持久化模块、需求模型模块、经典模型模块以及本体模块；最后，对于每个模块进行了详尽的设计，并给出不同模块之间的协同关系，完成了对信息检索模型的综合研究。

第五章 基于工程测试技术本体的语义检索系统

基于第三章已经构建好的工程测试技术本体和第四章的语义检索模型，本章实现了基于工程测试技术本体的语义检索系统。首先进行系统的需求分析，这是整个系统的业务逻辑基础；然后采用第五章提出的语义检索模型进行语义扩展，并对第三章生成的本体 OWL 文件进行解析，随后构建出基于工程测试技术本体的语义检索系统；最后对系统进行实验，结果充分说明本系统具有检索的语义性的特点，提高了信息检索的智能性，同时也说明了本文针对工程测试技术领域的企业级知识管理应用提出的这套解决方案是行之有效的。

5.1 系统需求分析

在工程测试技术领域，随着技术的更新以及经验的积累，对该领域的知识管理需求越来越大，对该领域的技术、经验知识等的学习与传承更是技术发展的重中之重。这些需求的一个重要的形式就是针对该领域的信息检索，而信息检索的流程中，如果仅仅依靠关键字匹配的方式来返回对应的资源，显然无法做到精准的语义查询，这种方式只有资源与用户发出的查询指令完全一致或者字符串相似度较高，用户才会获得相应的结果，不能很快的帮助工程测试人员建立起对该领域技术的认识。于是，本文将已经构建完成的工程测试技术本体以及提出的语义检索模型进行应用层面的实现，开发出了基于工程测试技术本体的语义检索系统。

本文提出的基于工程测试技术本体的语义检索系统主要功能涵盖以下四个方面：

（1）查询代理模块对用户输入的查询请求进行预处理；

（2）利用用户需求模块将用户多次检索的需求形成需求模型，形成用户检索的扩展规则，产生更多的但是有限的特征词扩展规则，使系统能够更加充分地理解用户的查询意图；

（3）通过本体的语义推理功能，借助 Jena 推理机以及对用户需求模型的迭代，基于本体的工程测试技术检索原型系统对用户提出的查询指令进行基于语义的推理以真正得到用户的查询意图；

（4）最后，将检索结果大于一定阈值的文本知识按照其与查询指令的相关

性从大到小排列，展示给用户，完成语义检索。

5.2 系统介绍与平台选择

本节主要从系统介绍以及开发平台的选择两方面进行描述。

5.2.1 系统介绍

目前主流的系统架构主要有 C/S 和 B/S 架构，两者存在多方面的区别，下面对两者架构的特点作一些说明：

C/S 架构即通过客户端与服务器的连接模式。客户端为在用户主机上运行的应用程序，需要用户下载安装对应的客户端，通过地址访问服务器的方式与本地或远程的服务端进行通信。C/S 架构对于信息安全的控制能力较强，能发挥 PC 的全部性能，所以响应速度快，但维护不便，需要用户下载安装更新的数据包。

B/S 架构是网页和服务器响应模式。在服务端安装有 mysql 等数据库的情况下，用户使用支持的浏览器即可通过网页访问数据库。B/S 架构支持跨平台，跨系统，无需特定硬件，更加开放，也因此对信息的安全控制能力较弱。其响应速度相比 C/S 架构较慢，具有交互延迟，但维护方便，不需要用户下载安装更新的数据包。

综合两个架构的介绍和对比，系统选择 B/S 架构作为基础架构，客户端为用户设备中的浏览器，用户输入域名完成登录后即可使用本系统。相关数据信息存放在云服务器中的数据库，方便浏览器通过移动网络进行远程访问。B/S 架构保证了基于本体的工程测试技术检索原型系统的便于维护更新、可快速迭代的特性。基于本体的工程测试技术检索原型系统的架构，系统首先接收用户的查询请求，对该请求进行预处理，并对查询请求进行语义层面的扩展，依据推理规则进行语义推理，结合用户需求模型，得出完整的语义扩展集合，将检索结果排序输出。

5.2.2 系统开发平台及相关介绍

一、系统开发平台

本系统开发平台主要包括硬件开发平台和软件开发平台，硬件平台用于对系

统提供硬件的基本开发与运行的支持，软件平台主要用于给系统提供开发环境以及运行环境的支持。

1、硬件平台

（1）开发平台应用端：

本系统开发过程中需要工程测试技术领域本体进行推理和维护，因此对计算能力等硬件配置有一定要求。为了更流畅的开发本应用，推荐以下配置：

PC 处理器双核或以上，主频至少 2.8GHz；内存至少 8GB；硬盘保留 100GB 以上的可用存储空间，推荐系统盘使用固态硬盘；显卡 GT730 或以上，推荐 Quadro K1000 或以上，至少支持 Direct 9；

（2）开发平台数据端：

基于本体的工程测试技术检索原型系统中所需的数据主要存放在数据库中，为了便于移动设备访问，将数据放在服务器中，因此推荐使用云服务器，配置如下：

服务器处理器推荐 Intel Xeon E5-2682 v4；内存为 2 GB DDR4 或以上；硬盘 40GB 存储空间。

2、软件平台

基于本体的工程测试技术检索原型系统需要需要跨平台支持，因此选择 java 语言作为开发语言，集成开发环境选择 IntelliJ IDEA 2018。PC 平台采用 Windows 10 操作系统，数据库采用 Mysql 5.7，服务器采用 Tomcat 服务器。为更快捷的设计前端界面，本文采用 LayUI 框架进行辅助设计。

二、相关软件介绍

1、IntelliJ IDEA 2018

IntelliJ IDEA 2018，是 java 编程语言开发的集成环境，由 JetBrains 公司研发。IntelliJ IDEA 在业界被公认为最好的 java 开发工具，尤其在智能代码助手、代码自动提示、重构、JavaEE 支持、各类版本工具(git、svn 等)、JUnit、CVS 整合、代码分析、 创新的 GUI 设计等方面的功能可以说是超常的，它的旗舰版本还支持 HTML，CSS，PHP，MySQL，Python 等。它的优势包括：

（1）强大的整合能力。比如：Git、Maven、Spring 等；

（2）提示功能的快速、便捷；

（3）提示功能的范围广；

（4）好用的快捷键和代码模板。比如：private static final = psf；

（5）精准搜索。

2、MySQL 数据库

MySQL 数据库是一种多用户、多线程且开源的关系型数据库管理系统，相比其他数据库系统，其体积较小，速度快，可靠性高，适应性较强，对市面上几乎所有的操作系统具有较好的兼容性和支持。并且，MySQL 数据库易于获得，可以很方便的下载安装，新用户能够较快掌握其使用方法。

3、LayUI 框架

LayUI 是一款采用自身模块规范编写的前端 UI 框架，遵循原生 HTML/CSS/JS 的书写与组织形式，门槛极低，拿来即用。其主要提供了很多好看、方便的样式，并且基本拿来即用，和 Bootstrap 有些相似，但该框架有个极大的好处就是定义了很多前后端交互的样式接口，如分页表格，只需在前端配置好接口，后端则按照定义好的接口规则返回数据，即可完成页面的展示，极大减少了后端人员的开发成本。

5.3 系统功能实现

5.3.1 界面设计

一、界面设计原则

UI 界面是用户接触使用本系统的关键媒介，图形界面的设计不仅需要符合交互原理，也需要按照特定的交互流程以及规范进行合理的设计。因此，在基于本体的工程测试技术检索原型系统中按照以下原则进行设计。

（1）一致性原则

一致性原则的主要任务是给用户提供相对稳定的交互风格，进而带来更好的交互体验。该原则能够保证用户在使用该系统时的操作稳定性，保持良好的操作习惯，进而使用户对交互有更好的理解，从而提高效率。

（2）多样化原则

多样化原则即界面设计需要注重形式和内容的多样化。工程测试技术信息通

过更多形式的表达才能让用户产生利于交互的思维。除去传统的文字信息，引入匹配图片，最大限度呈现多样化。

（3）高效性原则

高效性即用户界面应保证用户交互的高效，减少繁琐不必要的操作。按钮、滑动条等交互控件旨在提高用户与系统的交互能力和效率。而控件的合理的布局对于交互也至关重要，应当尽可能减少用户交互所需的操作次数。

（4）功能性原则

功能性原则即用户界面的设计要以实现交互或信息语义检索为导向，将其贯穿设计的始终。UI 界面的最终目的是辅助用户进行交互，保证用户在交互上能够获得最大程度的便利。

二、UI 界面的开发方法

UI 界面是在 IntelliJ IDEA 软件中完成开发，采用 HTML5 来设计界面，并使用 LayUI 框架对界面进行修饰。图 5.1 是对前端界面开发流程的展示：

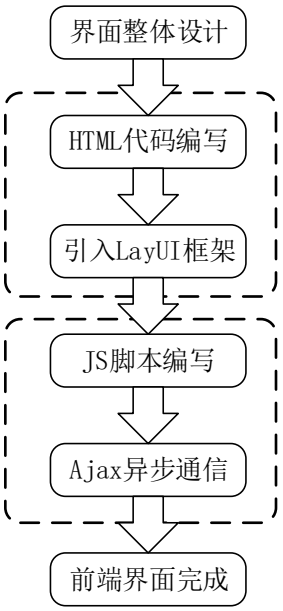

图 5.1 界面开发流程

1、界面整体设计

本文将基于上述界面设计原则，首先对界面进行设计，主要界面包括：登录界面、主界面、布尔检索界面、语义检索界面、文件上传界面、文件下载界面等。为增强用户使用体验的一致性，本文将每个界面都存在相同的主菜单界面，对应的界面选项高亮显示。

2、HTML 代码编写及引入 LayUI 框架

HTML 是设计界面的第一要素，本文将使用 HTML5 来进行静态页面的编写，并在其中引入对 LayUI 框架的依赖，将 js 脚本以及 Ajax 脚本也嵌入其中，使得界面具有动态性。

3、JS 脚本编写及 Ajax 异步通信

JS 是在浏览器下运行的脚本语言，能够与后台进行交互，具备动态刷新界面的能力，而 Ajax 则可以满足界面的局部刷新的要求，因此，本文使用 JS 与 Ajax 结合的界面开发方式，使系统界面更加简洁，响应速度更快。图 5.2 是在开发基于工程测试技术本体的语义检索系统时的部分前端代码：

```
const del = function () {
    $('.del').click(function () {
        const id = $(this).parents('tr:first').data('id');
        //删除工程测试技术本体检索系统数据库中的文件
        layer.confirm('您确定要删除本文件吗？', {btn: ['确定', '取消'], icon: 3}, function () {
            ajax({
                url: "../api/delFile",
                data: {id: id},
                success: function (res) {
                    if (res.code > 0) {
                        layer.closeAll();
                        layer.msg("删除成功！", {icon: 1});
                        getFileList();
                    } else {
                        layer.msg("删除失败，请重试！", {icon: 2});
                    }
                }
```

图 5.2 前端界面设计部分代码

5.3.2 Jena 推理机应用

进行语义推理的前提是合适的的推理机。Jena 是由惠普实验室开发的一套可以通过自行规定的函数直接调用程序功能的语义推理工具，Jena 是面向语义 Web 的应用开发包，包含的内容比较全面，推理机只是其中一部分。Jena 提供的推理机也和 RACER、FaCT、Pellet 等一样，是针对本体的推理机，并且其支持对 OWL 本体描述语言的解析和推理。因此，检索模型的本体模块，本文采用 Jena 推理机，主要使用 Jena 的两大功能：解析 OWL 形式化语言和本体推理。

一、解析 OWL 形式化语言

Jena 具有内置的 Ontology Api，基于此可以进行 OWL 文件的读入，并建立 Model 对象，然后，对 OWL 结构的节点进行一步步的解析，从根节点开始向下，

逐步迭代，直至将所有非叶子节点和叶子节点全部遍历。最后，将解析出的本体结构映射到 Model 对象内。

二、本体推理

本体推理需要先注册推理机，然后对上一步骤解析得到的 Model 对象进行推理。例如，A 是 B 的父节点，B 和 C 是兄弟节点，那么可以推理出 A 是 C 的父节点。

5.3.3 主要逻辑实现

系统的主要逻辑分为技术逻辑和功能逻辑。技术逻辑是指不同技术之间如何进行衔接合作，功能逻辑是为了实现功能需求而设置的逻辑。

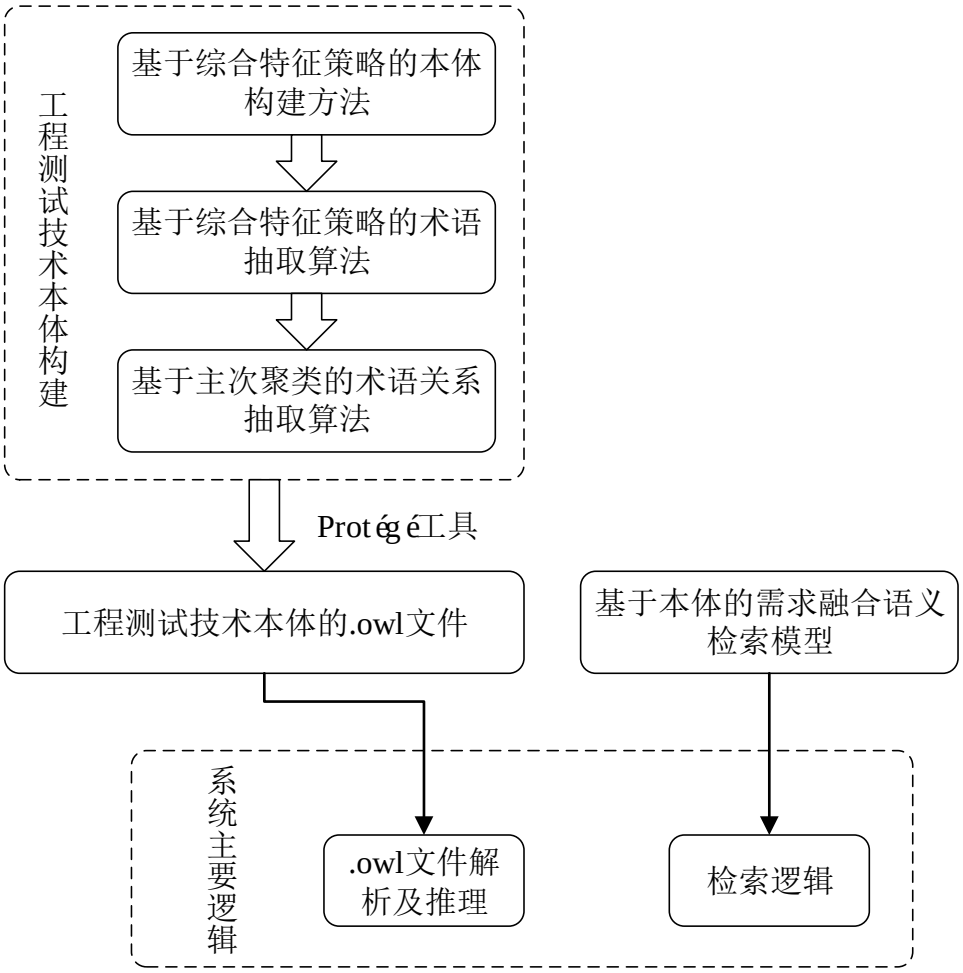

图 5.3 系统实现的技术逻辑

系统实现的技术逻辑如图 5.3 所示：在第二章提出的术语及术语关系抽取算法作用下，从文本中抽取出了工程测试技术领域的术语和术语关系，protégé 工

具将这些术语及术语关系加以编辑生成该领域本体的.owl 文件。系统的技术逻辑是：首先在磁盘中读取工程测试技术本体的.owl 文件，进行解析推理。其次，实现语义检索模型的检索逻辑。最后，整合这两个模块完成系统的技术逻辑实现。

系统的功能逻辑包括用户输入分解以及本体推理设计：

一、用户输入分解

用户在进行查询行为时，有时会输入空格隔开的关键词集合，有时会输入一个非结构化的句子。为了将对检索系统的输入统一为关键词集合，本文将对第二种用户输入的句子进行分词处理，得到初始特征词集合。本文基于正向最大匹配算法进行分词，算法流程如图 5.4 所示：

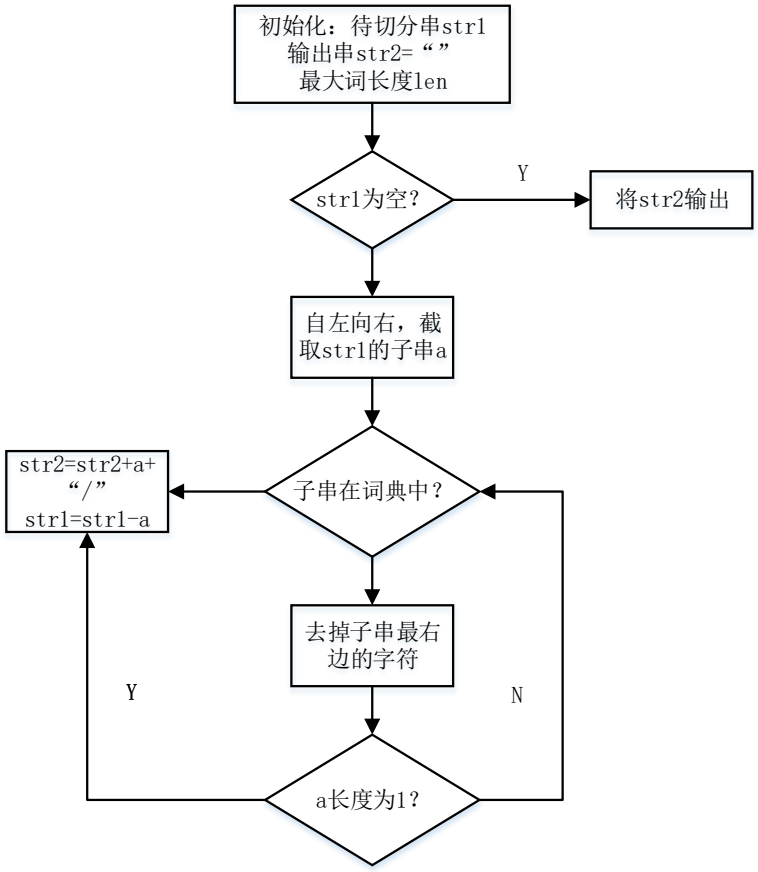

图 5.4 正向最大匹配分词算法流程图

在经过分词之后，用户输入就统一成关键词集合，为后续检索提供基础。系统将使用 java 中的 ArrayList 集合存储关键词集合。

二、本体推理设计

本体推理功能主要负责将用户提供的关键词进行扩展。本文设计了基于 Jena 推理机的方式进行本体推理。

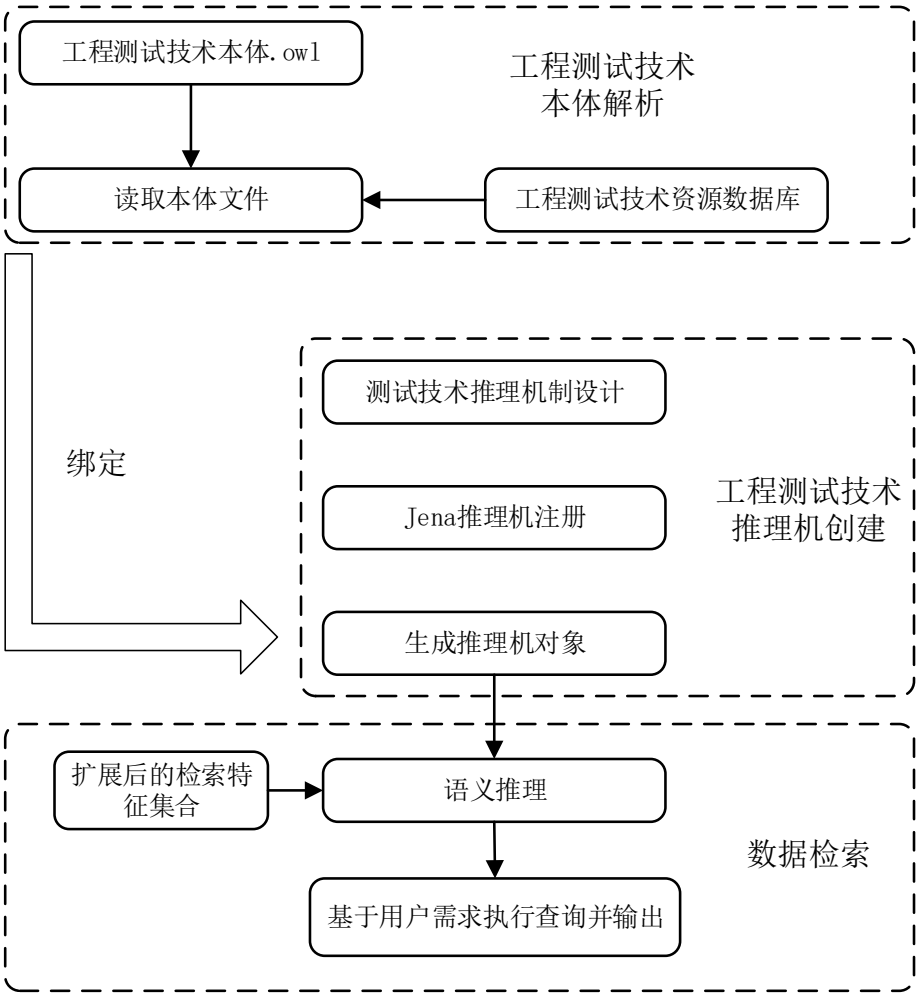

图 5.5 语义检索功能逻辑

基于工程测试技术本体的语义检索系统的语义检索功能逻辑如图 5.5 所示：分为三个模块：工程测试技术本体解析、工程测试技术推理机创建以及数据检索。

工程测试技术本体解析模块主要进行本体文件的读取，和解析。由于本文的本体持久化设计为文件管理与数据库结合的方式，因此，此处可以直接通过 IO 流的方式将工程测试技术本体的 OWL 文件进行读取操作并解析到内存上。

工程测试技术推理机创建模块主要负责对 Jena 推理机的注册。首先，在 jvm 内存上生成一个推理机对象，其次，定义推理规则，例如，根据工程测试技术本体信息的特点对规则做出定义，测试工具 X 是 Y 的一种，Y 具有功能 Z，则工具 X 具有功能 Z，其规则在 Jena 推理机中的表述如下：

Rule1：(?X Kind of ?Y) ，（Y Has Function of ?Z）>（X Has Attribute of ?Z）。

将定义的规则、本体解析过程的 Model 对象与 T 推理机对象进行绑定，得到了工程测试技术本体推理模型。之后，将结合用户需求模型，得到用户需求的

语义关系。例如，某用户常用“应用于”关系和“辅助”关系，那么将具有此关系的类别下的实例返回给用户。

语义检索过程中本体推理的主要逻辑的部分执行代码如图 5.6 所示：

```
//step1:解析 owl，生成 Model 对象（model）
// step2:编写 String 类的 SPARQL 语句（userString）并创建 query 对象
 Query query=QueryFactory.create(userString);
 //step4:创建查询执行对象
 QueryExecution queryExecution=QueryExecution.create(query,model);
 //step5:执行查询，获取结果集
 ResultSet resultSet=new queryExecution.execSelect();
 //step6:解析结果集
 ResultSetFormatter.out(System.out,resultSet,query);
```

图 5.6 本体推理模块部分逻辑代码

5.3.4 系统主要界面展示

本节主要展示原型系统的登录界面、主界面、布尔检索界面、语义检索界面以及文件上传界面等五个界面，并且对每一个界面进行介绍。

一、登录界面展示

图 5.7 系统登录界面

图 5.7 展示了基于工程测试技术本体的语义检索系统的登录界面，用于用户登录，其内容包括用户名和密码输入以及登录按钮。系统将在后台判断用户身份：普通用户、系统管理员以及数据管理员。

二、首页界面展示

图 5.8 系统首页界面

图 5.8 展示了基于工程测试技术本体的语义检索系统的首页，主要包含三个功能区间：左侧的主菜单、上层的用户登录信息以及中间的检索结果显示。其中，主菜单包括：布尔检索、语义检索、文本数据添加、数据导出、检索记录索引以及需求设置等功能。上层用户登录信息功能区间包括：当前时间、管理员信息管理、当前用户以及退出按钮。中间的结果展示区则用于进行数据的展示。

三、布尔检索界面展示

图 5.9 布尔检索界面

图 5.9 展示了系统的布尔检索界面。布尔检索是本文为满足用户的简单检索需求而开发的功能，用户在检索需要的文本数据时，系统只根据文件名的关键词进行匹配。对于结果的展示将通过表格的方式，将系统返回的结果的文件名、上一主题、大类以及文件位置、文本创建时间以及操作作为表格的字段，用户查看某一文件只需点击“查看”按钮，同样，需要编辑和删除时单击对应的功能按钮即可。从图中可以看出，在搜索 “传感器”时，只返回了力传感器和位移传感器等文件名包含“传感器”的文件。

四、语义检索界面展示

图 5.10 展示了系统的语义检索界面。该界面的功能以及交互逻辑如下：首先，在搜索框输入检索关键词；其次，选择需求融合、父类扩展以及子类扩展的具体选项；最后，点击语义检索按钮，得到搜索结果。本文以“力传感器”为检索关键词进行了语义检索功能的结果展示，从图中可以看出，系统不仅返回了力传感器的文本，也将力传感器的理论基础以及相关传感器进行了结果的返回。如果用户不需要语义的扩展，那么可以使用布尔检索或者修改需求融合的选择。

图 5.10 语义检索界面

五、文件上传界面展示

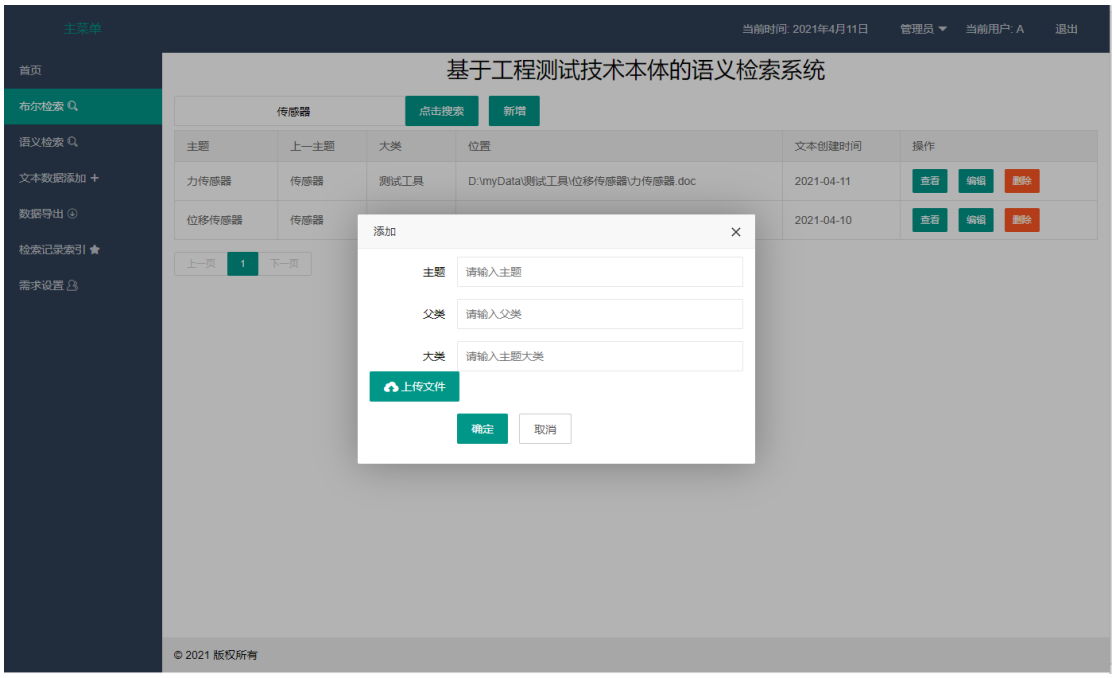

图 5.11 文件上传界面

图 5.11 展示了系统的文件上传界面。该界面的功能以及交互逻辑如下：首先，在文件显示的功能区上方点击新增按钮；其次，在弹出的文件上传信息框中输入文件的主题、父类以及大类；最后，通过上传文件按钮的文件上传功能选择对应的文件，即可将文本文件添加到系统中。

5.4 系统测试与分析

5.4.1 系统测试

语义检索模型的性能和基于综合特征策略方法构建的本体的质量可以由用户对检索结果的满意程度反映出来，检索返回的数据集合越能够满足用户的检索意图，则检索的质量越高，本体质量以及检索模型的性能也越好，反之则反。本文选用多次信息检索的查准率和查全率两个性能指标来评价检索模型的性能和本体质量的评价。

一、系统测试准备

针对基于工程测试技术本体的语义检索系统的测试主要包括以下准备：

（1）检索指令的准备

检索指令是指用户的检索语句或者检索关键词。本文的检索需求分别覆盖工

程测试技术领域的理论、测试目的、测试方法、测试意义、测试经验、测试工具以及应用方式等七个大类。本文对每一个大类准备了100条检索指令，共700条。

（2）检索需求的准备

检索需求包括：关键词检索需求、语义检索需求、用户需求融合检索需求，本文将检索指令按照以上需求平均分为三部分。

关键词检索即满足每一条包含关键词的数据都被返回即可；语义检索需求则以及用户之前检索的内容自动扩展语义，返回数据集合；用户需求融合检索需求即根据用户手动选择的需求，例如，扩展父类或者扩展子类等需求，将对应的数据集合加以返回；

（3）检索参考结果的准备

由于检索的结果针对每次检索是不同的，因此，本文将依据检索需求和数据库中的文件内容，给出参考的返回数据集合。例如，现有一条语义检索指令，要求其具有用户需求融合的功能，主要扩展某特征节点的子类和它的兄弟节点，那么，本文将其子类和所有兄弟节点作为参考的返回数据集合。这样设计的好处是可以最大程度的避免供参考的数据集合的定义的主观性，根据需求来定义正确的结果，这种方式是非常客观可信的。

二、系统测试流程

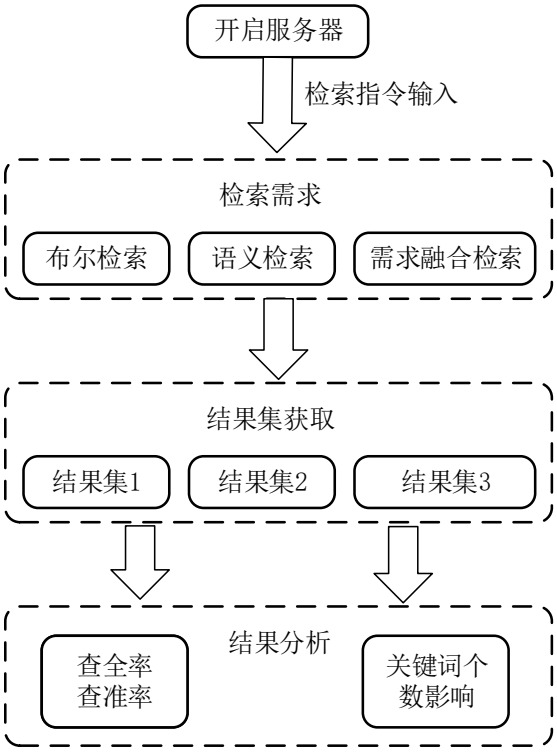

图 5.12 系统测试流程

基于工程测试技术本体的语义检索系统的测试如图 5.12 所示，包括：开启服务器、检索需求、结果集获取、结果分析等四个部分，以下是对四个步骤的详细描述：

（1）开启服务器

本文使用了 IntelliJ IDEA 的整合 tomcat 服务器的功能，首先将项目直接通过此开发环境工具来部署到 tomcat 服务器上，其次，确定项目的端口号等信息，最后开启服务器。

（2）结果集获取

结果集获取主要是将检索结果进行整理。首先，根据三类不同的检索需求将检索结果存储在三个不同的位置；其次，将预定义的参考结果集也按照此分类方式进行分类；最后，将不同检索需求下的真实结果集与参考结果集存储在同一位置。

（3）结果分析

首先，将每个结果集对应的检索指令与结果匹配预先定义的结果集，针对每一次检索，计算查准率与查全率；然后，将查全率与查准率进行汇总，整体分析系统的性能；最后，将三类检索需求统一分析，得出关键词个数对于检索效果的影响。

5.4.2 测试结果分析

一、查全率查准率分析

通过上述实验，本文得出了不同检索方法下的检索效果，量化指标通过查全率/查准率以及 F 值进行衡量。三种方法的关系为：关键字检索只依靠关键词匹配进行查询，而语义检索是基于对本体的推理进行查询，需求融合检索是在语义检索的基础上融合了用户的需求，也就是基于本文提出的需求融合模型进行的检索。

从图 5.13 可以得出：关键词检索方法的查准率和查全率都相对另外两种方法低；语义检索查全率与需求融合的查全率非常接近，但是查准率相对于需求融合的检索方法就明显降低，这是因为，不针对用户需求的语义扩展为了提升查全率返回了过多的数据，这其中就包含用户不需要的数据。因此，可以验证本文提

出的基于本体的信息检索是有效的，比关键词检索的 F 值提高了 25.7%；并且加入了需求融合处理后的检索方法也是有效的，比基于本体的语义检索 F 值提高了 5.2%，即本文提出的检索方法较关键词检索的 F 值提升了 30.9%。

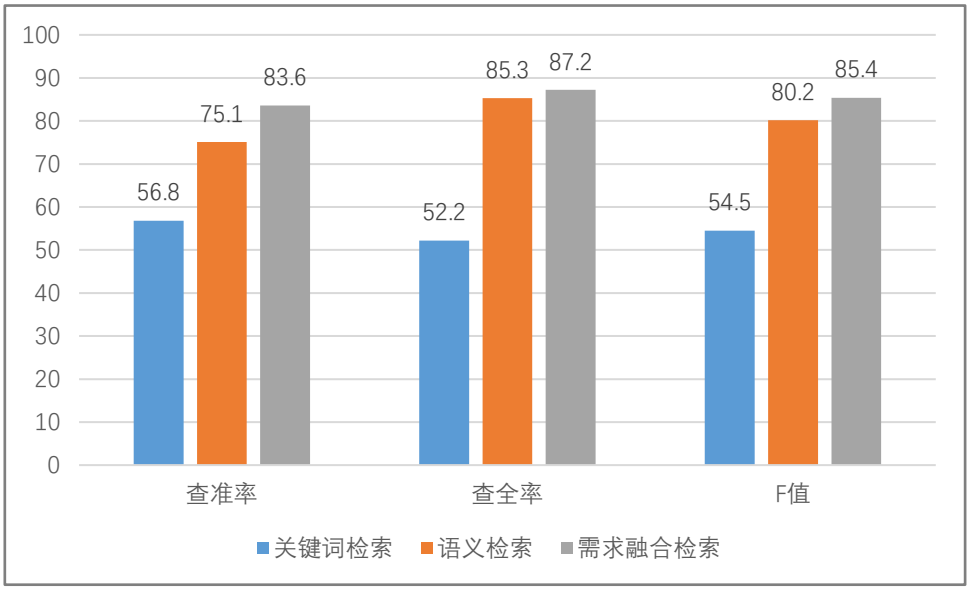

图 5.13 不同检索方法的性能

二、关键词个数分析

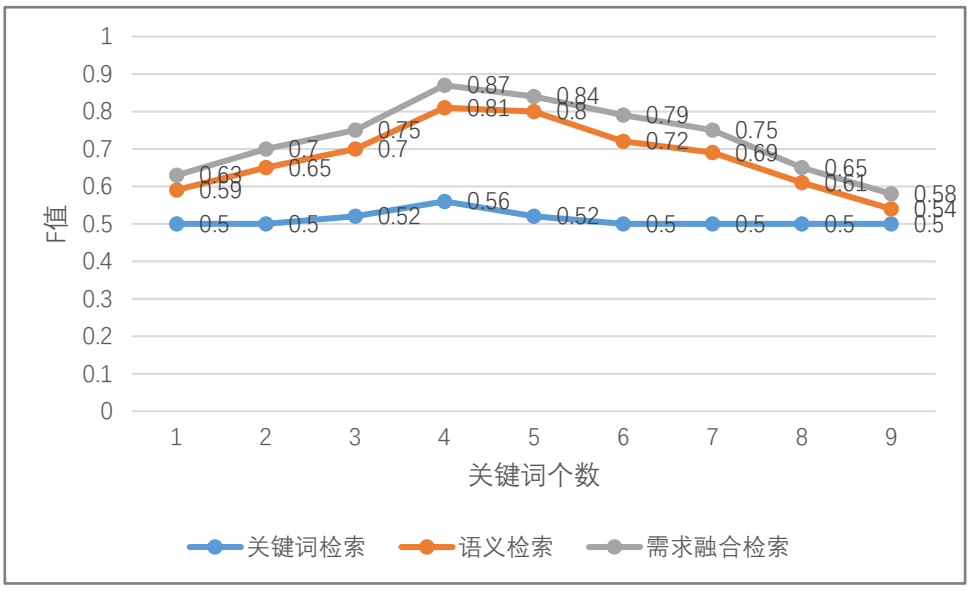

图 5.14 关键词个数对检索 F 值的影响

从图 5.14 可以得出，关键词检索、语义检索以及基于需求融合的语义检索三种方法，随着检索关键词的个数的增加，检索结果的 F 值呈现先上升后下降的趋势。经过分析，折线上升的原因是在关键词个数小于 4 时，随着关键词个数的增加，系统能够更好的扩展语义，因此 F 值呈现上升的趋势；当关键词个数大于 4 时，由于本系统在语义扩展时，若语义扩展较多将不利于用户需求内容的发现，

只会选择其中一种重要的扩展关系进行语义扩展，例如，父子类关系，因此随着关键词个数的增加，检索的 F 值逐渐下降。总之，使用本系统进行检索时关键词个数最好选择 4，此时检索效果的 F 值最大。

5.5 本章小结

本章主要进行了基于工程测试技术本体的语义检索原型系统的设计与开发。首先，明确了基于工程测试技术本体的语义检索原型系统的核心需求：语义检索以及加上需求融合的语义检索。其次，对系统的开发平台以及相关工具进行了介绍。然后，对系统进行功能的实现，主要包括：界面设计、Jena 推理机应用、主要逻辑实现以及界面展示。最后，对系统进行了测试，在分析测试结果的同时验证了本文提出的本体抽取算法以及语义检索模型的优越性。

第六章 结论与展望

6.1 结论

工程测试技术是观察产品质量和维护产品性能最重要的技术。工程测试技术是我们从事生产，进行科学研究的重要手段，是国民经济发展和社会进步所需的必不可少的技术。工程测试技术的不断更新，给该技术的学习、使用以及传承带来了极大的挑战。因此，结合本体半自动构建技术以及信息检索理论的相关研究，设计开发基于工程测试技术本体的语义检索系统，对于提高技术学习的效率，帮助企业测试技术以及经验的传承具有极大的价值。

本文首先对课题的研究背景，意义以及关键技术的国内外研究现状进行调查和阐述；其次改进了本体抽取的两个算法；然后提出了基于综合特征策略的本体构建方法并进行工程测试技术本体的构建；接着设计了基于本体的需求融合语义检索模型；最后，完成了基于工程测试技术本体的语义检索原型系统的设计与实现，具体内容如下：

1、介绍了课题的目的，意义，以及相关理论技术的国内外研究现状。主要包括术语（关系）抽取，信息检索模型，本体构建方法的研究现状。

2、进行了术语抽取及术语关系抽取算法的改进。本体抽取包括术语抽取以及术语关系抽取两个部分。因此，本文将本体抽取算法分两部分处理：基于综合特征策略的术语抽取算法和基于主次聚类的术语关系抽取算法。基于综合特征策略的术语抽取算法对于从文本中抽取术语的过程中产生的不同特征进行了分析，针对不同的特征提出了具体的策略。基于主次聚类的术语关系抽取算法主要对K-means 聚类算法进行了改进，并使用两次 K-means 聚类结合特征提取的方式抽取术语关系，为下一步本体构建提供算法基础。

3、构建了工程测试技术信息本体。首先，对工程测试技术领域本体构建的主要原则进行分析并提出基于综合特征策略的本体构建方法；然后，按照本文提出的方法，对工程测试技术领域本体进行构建；最终，得到了工程测试技术领域本体，并设计了该领域本体的持久化方法，生成了本体的 OWL 文件。

4、设计了基于本体的需求融合语义检索模型。本章首先通过分析目前信息

检索的趋势，提出了用户需求模型。然后，提出了基于用户需求模型的语义检索模型。最后给出了不同模块的主要内容以及他们之间的逻辑关系。

5、基于工程测试技术本体的语义检索系统设计与实现。基础是上面产生的本体的 OWL 文件和检索模型。首先，介绍了系统的软硬件平台和要求。然后，分析系统的需求功能，详细说明了本体推理和信息检索系统的开发技术。接着开发完成，展示了主要界面。最后，对系统进行了测试，并进行结果分析。

本文主要创新点：

1、对本体半自动构建方法进行研究，设计出了基于综合特征策略的本体构建方法，减少了本体构建的对人工的依赖，缩短了本体更新的周期。

2、提出了改进的本体抽取算法，主要包括术语抽取算法和术语关系抽取算法。改进的术语抽取算法通过针对不同的特征给出不同的策略进行细粒度的处理。术语关系抽取使用主次聚类与特征选择结合的方法进行关系抽取，并提出了基于特征的变密度 K-means 聚类算法，有效的提高了本体抽取算法的准确率。

3、提出了基于本体的需求融合语义检索模型。分析目前信息检索的趋势，提出了用户需求模型，并通过将传统检索模型、本体和用户需求模型的整合提出了基于本体的需求融合语义检索模型，该模型可以满足用户信息检索的语义性以及个性化检索，提高了信息检索的准确率。

6.2 展望

本文所研究的内容涉及的领域和范围较广，基于工程测试技术本体的语义检索系统还可以从以下角度进行深入研究：

1、基于综合特征策略的本体构建方法属于半自动本体构建方法，还无法完全脱离对人工操作的依赖，后续需要对本体自动构建方法进行研究。

2、目前只是对数据库中的文本数据进行知识管理，文件类型略显单一，如何将丰富的数据类型加以管理，使得用户或者企业在使用本系统时有更好的沉浸式体验，是后续需要研究的一个方向。

3、本文提出的术语抽取和术语关系抽取算法只能对中文文本进行抽取，不能应对多语种的文本，如何能从多语种文本中抽取本体也是后续需要研究的一个方向。

参考文献

【1】. 刘红. 大数据的本体论探讨[J].自然辩证法通讯,2014,36(06):115-121+128.

【2】. 李雪迪. 基于本体论的精细化数据分析[D].南京:南京邮电大学,2015.

【3】. 刘文韬,陈智宏,许焱,李星毅.基于本体论的交通异构数据集成系统[J].计算机系统应用,2010,19(03):7-11.

【4】. 郑亮. 应用本体论构建数据挖掘知识管理系统[D].重庆:重庆大学,2008.

【5】. 叶鹰,金更达.基于元数据的信息组织与基于本体论的知识组织[J].大学图书馆学报,2004(04):43-47.

【6】. 马路遥,夏博,肖叶,荀恩东.面向句法结构的文本检索方法研究[J].电子学报,2020,48(05):833-839.

【7】. 李宇,刘波.文档检索中文本片段化机制的研究[J].计算机科学与探索,2020,14(04):578-589.

【8】. 黄丽娟,周海.基于情感分析的文本检索系统的研究[J].科技创新与应用,2019(36):58-59.

【9】. 马天松.基于本体的企业管理领域数据知识表示方法分析[J].商场现代化,2020(17):120-122.

【10】.郑学伟.基于知识管理的本体自动构建算法研究[J].计算机技术与发展,2014,24(12):64-68.

【11】.马志斌. 特定领域术语自动抽取方法的研究[D].哈尔滨:哈尔滨工业大学,2009.

【12】.Song Xinyu,Feng Ao,Wang Weikuan,Gao Zhengjie,Zhang Feng. Multidimensional Self-Attention for Aspect Term Extraction and Biomedical Named Entity Recognition[J]. Mathematical Problems in Engineering ,2020.

【13】.Heng Yang,Biqing Zeng,Jianhao Yang,Youwei Song,Ruyang Xu. A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction[J]. Neurocomputing,2021,419.

【14】.Hua Ju,Shunwuritu Na. Research on term extraction technology in computer field based on wireless network technology[J]. Microprocessors and

Microsystems,2020(prepublish).

【15】. Rebekah Kennedy,Richard J. M. Reardon,Oliver James,Cherith Wilson,Padraic M. Dixon. A long‐term study of equine cheek teeth post‐extraction complications: 428 cheek teeth (2004‐2018)[J]. Equine Veterinary Journal,2020,52(6).

【16】. 闫琪琪,张海军.一种混合策略的领域术语自动抽取方法[J].电子制作,2015(08):50-51.

【17】. 梁颖红,张文静,周德富.基于混合策略的高精度长术语自动抽取[J].中文信息学报,2009,23(06):26-30.

【18】. 刘辉,刘耀.基于条件随机场的专利术语抽取[J].数字图书馆论坛,2014(12):46-49.

【19】. 古迎志. 基于术语抽取与匹配的推送技术及应用[D].北京工业大学,2018.

【20】. 俞琰,陈磊,姜金德,赵乃瑄.基于依存句法分析的中文专利候选术语选取研究[J].图书情报工作,2019,63(18):109-118.

【21】. Azanzi Jiomekong,Gaoussou Camara,Maurice Tchuente. Extracting ontological knowledge from Java source code using Hidden Markov Models[J]. Open Computer Science,2019,9(1).

【22】. Wen Zeng, Hongjiao Xu, Junsheng Zhang. Term extraction and correlation analysis based on massive scientific and technical literature[J]. Int. J. of Computational Science and Engineering,2017,15(3/4).

【23】. 董洋溢,李伟华,于会.文本特征和复合统计量的领域术语抽取方法[J].西北工业大学学报,2017,35(04):729-735.

【24】. 刘里,肖迎元.基于术语长度和语法特征的统计领域术语抽取[J].哈尔滨工程大学学报,2017,38(09):1437-1443.

【25】. 朱惠,王昊,苏新宁,邓三鸿.汉语领域术语非分类关系抽取方法研究[J].情报学报,2018,37(12):1193-1203.

【26】. Jiang Zhiying,Gao Bo,He Yanlin,Han Yongming,Doyle Paul,Zhu Qunxiong,Zeng Nianyin. Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports[J]. Mathematical

Problems in Engineering,2021,2021.

【27】. Yu Hailong,Ji Yannan,Li Qinglin,Wagner Neal,Sundhararajan ,Son Le Hoang,Joo Meng. Student sentiment classification model based on GRU neural network and TF-IDF algorithm[J]. Journal of Intelligent & Fuzzy Systems,2021,40(2).

【28】. 韩红旗,安小米.C-value 值和 unithood 指标结合的中文科技术语抽取[J].图书情报工作,2012,56(19):85-89.

【29】. 王均玲.大数据分析技术的数字图书馆信息检索模型设计[J].现代电子技术,2020,43(17):155-157+161.

【30】. 龚庆雄. 基于实体的信息检索模型研究[D].武汉:华中师范大学,2020.

【31】. C. Nagarjuna,D. Khadar Hussain,S. Vasundra. Extended Boolean Retrieval Model using P-Norm and Term Independent Bound Methods[J]. International Journal of Management, IT and Engineering,2014,4(7).

【32】. Daniel Z. Zanger. Interpolation of the extended Boolean retrieval model[J]. Information Processing and Management,2002,38(6).

【33】. Donald B. Cleveland,Ana D. Cleveland,Olga B. Wise. Less than full-text indexing using a non-boolean searching model[J]. Journal of the American Society for Information Science,1984,35(1).

【34】. 马艳荣,温煜坤.基于向量空间模型的对外汉语应用文写作词汇分类系统研究[J].现代电子技术,2021,44(08):137-140.

【35】. 冀晓玲. 基于概率模型的离群点检测近似算法的研究与实现[D].沈阳:沈阳航空航天大学,2019.

【36】. 张雪娜. 基于文档检索和语义关系识别的石油领域本体自动化构建[D].北京:中国石油大学,2017.

【37】. 唐飞. 生物医学领域本体自动构建方法的研究与实现[D].北京:北京理工大学,2016.

【38】. 段炼. 基于文本分析的石油领域本体自动构建方法的研究[D].大庆:东北石油大学,2015.

【39】. 杨靖. 领域本体自动构建的关键技术研究[D].哈尔滨:哈尔滨工业大

学,2008.

【40】. Kai Xu,Zhenguo Yang,Peipei Kang,Qi Wang,Wenyin Liu. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition[J]. Computers in Biology and Medicine,2019,108.

【41】. Yang Xi,Bian Jiang,Hogan William R,Wu Yonghui. Clinical concept extraction using transformers.[J]. Journal of the American Medical Informatics Association : JAMIA,2020,27(12).

【42】. López Úbeda Pilar,Díaz Galiano Manuel Carlos,Martín Noguerol Teodoro,Luna Antonio,Ureña López L. Alfonso,Martín Valdivia M. Teresa. COVID-19 detection in radiological text reports integrating entity recognition[J]. Computers in Biology and Medicine,2020,127.

【43】. Guo Xuchao,Zhou Han,Su Jie,Hao Xia,Tang Zhan,Diao Lei,Li Lin. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism[J]. Computers and Electronics in Agriculture,2020,179.

【44】. Dewandaru Agung,Widyantoro Dwi Hendratmo,Akbar Saiful. Event Geoparser with Pseudo-Location Entity Identification and Numerical Argument Extraction Implementation and Evaluation in Indonesian News Domain[J]. ISPRS International Journal of Geo-Information,2020,9(12).

【45】. 张二艳. 术语自动抽取技术研究[D].哈尔滨:哈尔滨工业大学,2009.

【46】. 侯庆霖.基于词向量及术语关系抽取方法的文本分类方法[J].移动通信,2018,42(07):12-17+23.

【47】. 韩红旗,徐硕,桂婕,乔晓东,朱礼军,安小米.基于词形规则模板的术语层次关系抽取方法[J].情报学报,2013,32(07):708-715.

【48】. Dou Xinyu,Liao Cuijuan,Wang Hengqi,Huang Ying,Tu Ying,Huang Xiaomeng,Peng Yiran,Zhu Biqing,Tan Jianguang,Deng Zhu,Wu Nana,Sun Taochun,Ke Piyu,Liu Zhu. Estimates of daily ground-level NO2 concentrations in China based on Random Forest model integrated K-means[J]. Advances in Applied Energy,2021,2.

【49】. K. Aseem,S. Selva Kumar. Hybrid k-means Grasshopper Optimization Algorithm based FOPID controller with feed forward DC–DC converter for solar-wind generating system[J]. Journal of Ambient Intelligence and Humanized Computing,2021(prepublish).

【50】. Jing Jiankun,Ke Shizhen,Li Tianjiang,Wang Tian. Energy method of geophysical logging lithology based on K-means dynamic clustering analysis[J]. Environmental Technology & Innovation,2021(prepublish).

【51】. S. Vimal,Y. Harold Robinson,M. Kaliappan,K. Vijayalakshmi,Sanghyun Seo. A method of progression detection for glaucoma using K-means and the GLCM algorithm toward smart medical prediction[J]. The Journal of Supercomputing,2021(prepublish).

【52】. Wong Nathan,Kim Daehwan,Robinson Zachery,Huang Connie,Conboy Irina M. K-means quantization for a web-based open-source flow cytometry analysis platform.[J]. Scientific reports,2021,11(1).

【53】. 刘煜澄. 面向多源数据的军事本体构建系统[D].南京:东南大学,2019.

【54】. Sha Wang,Qiong Peng,Hua Liang. An Event Ontology Model Research for Environmental Pollution Emergencies[J]. International Journal of Ambient Computing and Intelligence (IJACI),2020,11(4).

【55】. Caihua Qiu. Ontology Mapping Constructing by means of Low Rank Distance Matrix Optimization[J]. Engineering Letters,2020,28(3).

【56】. Geng Qian,Deng Siyu,Jia Danping,Jin Jian. Cross-domain Ontology Construction and Alignment from Online Customer Product Reviews[J]. Information Sciences,2020(prepublish).

【57】. 童名文,牛琳,杨琳,邹军华,上超望.课程本体自动构建技术研究[J].计算机科学,2016,43(S2):108-112.

【58】. 廉龙颖.基于本体的网络空间安全知识图谱的构建方法[J].黑龙江科技大学学报,2021,31(02):254-258.

【59】. 尹弼民. 基于概念格理论的领域本体半自动构建方法研究[D].南昌:南昌大学,2015.

【60】. 尹峥晖. 基于叙词表的领域本体构建[D].长沙:湖南大学,2015.

【61】. Zhiguo Peng,Meifa Huang,Yanru Zhong,Zhemin Tang. Construction of ontology for auto-interpretable tolerance semantics in skin model[J]. Journal of Ambient Intelligence and Humanized Computing,2019(prepublish).

【62】. Qi Zhang,Yuanqiao Wen,Chunhui Zhou,Hai Long,Dong Han,Fan Zhang,Changshi Xiao. Construction of Knowledge Graphs for Maritime Dangerous Goods[J]. Sustainability,2019,11(10).

【63】. 舒慧欣. 基于关系数据库的领域本体构建方法[D].南昌:江西师范大学,2012.

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

【65】. 郭会雨. 疾病领域本体模型构建研究[D].北京:中国人民解放军军事医学科学院,2011.

【66】. 周炫余,唐祯,唐丽蓉,李璇,卢笑.基于多源异构数据融合的初中数学知识图谱构建[J].武汉大学学报(理学版),2021,67(02):118-126.

【67】. 刘宏义,杨明.基于DDS网络中间件的持久化数据管理问题研究[J].微电子学与计算机,2016,33(10):163-166+172.

【68】. 欧阳宏基,葛萌,陈伟.基于 JDBC 的数据持久化层性能优化研究[J].网络新媒体技术,2016,5(05):9-15.

【69】. 李苹,孙若贤.Hibernate 数据持久化方法应用研究[J].电脑知识与技术,2016,12(03):12-13.

【70】. 李松涛.基于 Kudu 的结构化数据存储方案设计分析[J].数字技术与应用,2019,37(10):183+185.

【71】. 于芳.基于用户需求驱动的个性化图书信息智能整合系统设计[J].现代电子技术,2020,43(11):158-162.

【72】. 蒋大平.优先聚合运算：用户需求信息检索的优化[J].绥化学院学报,2019,39(08):134-136.

【73】. 田巧萍,吕健,潘伟杰,王伟祎,袁涛.基于用户需求的产品深度个性化定制[J].图学学报,2018,39(05):867-878.

【74】. 安靖,陈宇行.形式化概念分析在信息检索中的应用[J].软件导

刊,2013,12(01):121-122.

【75】. Soni Sarvesh,Roberts Kirk. An evaluation of two commercial deep learning-based information retrieval systems for COVID-19 literature[J]. Journal of the American Medical Informatics Association,2021,28(1).

【76】. Tzu-Hao Lin,Yu Tsao. Source separation in ecoacoustics: a roadmap towards versatile soundscape information retrieval[J]. Remote Sensing in Ecology and Conservation,2020,6(3).

【77】. Twitter Inc.Patent Issued for Indexing Data In Information Retrieval Systems (USPTO 10,810,236)[J]. Internet Weekly News,2020.

【78】. L Arokia Jesu Prabhu,Sengan Sudhakar,G K Kamalam,J Vellingiri,Gopal Jagadeesh,Velayutham Priya,V Subramaniyaswamy. Medical information retrieval systems for e-Health care records using fuzzy based machine learning model[J]. Microprocessors and Microsystems,2020(prepublish).

【79】. 杨月华. 基于领域知识模型的突发事件智能信息检索系统研究[D].北京:北京邮电大学,2013.

【80】. 王进. 基于本体的语义信息检索研究[D].合肥:中国科学技术大学,2006.

在攻读硕士学位期间公开发表的论文

【1】. Xiaomei Hu, Wen Bo, Jianfei Chai. Research on the Visual Simulation Platform of Acupoint Massage Based on Unity 3D [C]. International Conference on Image, Vision and Computing, ICIVC2019, p1-5. (已被 EI 检索：20201908641626，导师第一作者，本人第二作者)

作者在攻读硕士学位期间所做的项目

【1】. 横向项目—水下声磁数字可视化软件

【2】. 横向项目—沉管隧道管段浮运、沉放对接定位系统监控软件技术开发

致　谢

本文是在导师胡小梅副研究员的悉心指导下完成的。承蒙胡老师的亲切关怀和精心指导，在攻读硕士研究生两年多的时间里，胡老师即使工作十分繁忙，仍然抽出宝贵的时间，不遗余力地给予我学术上的科学指导和细心的帮助，从论文选题、框架搭建、内容把握整个过程中所做出的巨大付出，使我从中获益不浅。胡老师对学生认真负责的态度、严谨的科学研究方法、敏锐的学术洞察力、勤勉的工作作风以及勇于创新、勇于开拓的精神是我永远学习的榜样。在此，谨向胡老师致以深深的敬意和由衷的感谢。

感谢在读期间实验室师兄弟的帮助，他们是徐慧靖师兄、李明杭师兄、张贻启师兄、缪佳舒师兄、潘兆仁师兄、徐俊师兄、王川师兄、吕顺可同学以及研二、研一的师弟师妹们，谢谢你们。

感谢我的室友何文韬、谌稳帅、乐宇倚在学校生活上的支持与帮助。

还要感谢我的父母和姐姐们，他们在生活上给予我很大的支持和鼓励，是他们给予我努力学习的信心和力量。

最后，感谢所有关心我、支持我和帮助过我的同学、朋友、老师和亲人。在这里，千言万语用一句话来表达：感谢你们，感恩你们！