摘 要

随着互联网上信息量的迅猛增加，语言学工作者研究的不断深入，如何在纷繁复杂的文献材料中，快速、准确地找到用户需要的信息，文本分类起着非常重要的作用。而其中，基于语义的文本分类逐渐成为主流，语义关系的最佳载体—本体，成为了学术界关注的焦点。

本体就是对概念和关系的描述，基于本体的文本分类就是基于知识层面和语义层次上的分类。本文在论述语言学领域本体库建立并应用于文本分类意义的基础上，通过细致分析语言学内部词语之间的语义关系，构建了语言学文献的领域本体。提出了基于本体语义关系进行匹配的语言学文献分类方法。首先，利用已有的分词系统对文献进行分词处理和关键词抽取，采用经过一定改进的 TFIDF 算法，对文献关键词进行特征选择，确定待分类文本的特征项。然后将这些特征项与领域本体库中存储的领域特征项进行匹配，从而得到文本的类别。本文介绍了两种分类算法，一种是基于概念语义的匹配，一种是非一致性模糊匹配。无论采取哪种算法，都可以在一定程度上弥补当前分类系统缺乏语义联系的不足，提高文本分类的准确性。

关键词 文本分类 本体 领域本体 语言学文献 匹配

Abstract

With the rapid increase of internet information and the linguist's lucubrating, text classification plays an important role in how we can scan and use the required information concerning Linguistics literature promptly. But among, the document categorization based on semanteme gradually becomes the mainstream, The semantic relations best carrier — Ontology, become attention focus in the academic.

Ontology is a description between the conception and the relation. The document categorization based on ontology is based upon the level of knowledge and semantic relations categorization indeed. This paper which is based upon the discussion of Linguistics document Feature-Database Establishment and the application on text classification Structures the Linguistics document Feature-database by analysing the internal relations of linguistics words Semantic Earnestly, and proposes Linguistics document classification method which is based upon Semantic relations match. Firstly, Use the participle system to choose the key word with the document. Use the TFIDF algorithm in feature Extraction with the key words to analyse feature item. Then the feature item Matches with the Feature-Database, Thus obtains the document's category. This paper designs two kinds of categorization methods. One is based on the concept semantics match and the other is uniformity fuzzy match. No matter what algorithm is selected, it can make up insufficient of current categorization deficient semantic relation to some extent. Enhance the document classification accuracy.

Key words: Document Categorization; Ontology; Domain Ontology; Linguistics document; Matching

河北大学
学位论文原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知， 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经发表或撰写的研究成果，也不包含为获得河北大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的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致谢。

作者签名：曹亚妹 日期：2007年6月6日

学位论文使用授权声明

本人完全了解河北大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，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，可以采用影印、缩印或其他复制手段保存论文。

本学位论文属于

1、保密 □ ，在_____年_____月_____日解密后适用本授权声明。

2、不保密 □ 。

（ 请在以上相应方格内打“√” ）

作者签名：曹亚妹 日期：2007年6月6日

导师签名：田学东 日期：2007年6月6日

第1章 引 言

1.1 研究背景及意义

在网络逐渐普及，并进一步成为人们主要信息查询手段的今天，很多信息查询不确切的问题经常困扰我们。当用户进行信息查询时，与要求不大相关的信息会大量涌现，如何在纷繁复杂的信息中找到自己想要的内容，成为信息获取首先需要解决的问题。为了能获得更多有价值的信息，提高对信息组织、整理的效率，必须对文本进行自动分类。信息检索和文本分类是两个相辅相成的过程。只有把握好“分类”这一关键环节，对信息的利用才能达到更满意的效果。

文本自动分类（Automatic Document Categorization）[1]就是利用计算机对文本集（或其他实体对象）按照一定的分类体系或标准进行类别划分。自动分类技术是有效运用信息的基础，是代替繁杂的传统人工分类方法的有效手段和必然趋势。利用先进的计算机技术和人工智能技术进行文本分类，不仅方便快捷，容易实现，节省大量的人力、物力，而且还可以进行更深层次的信息挖掘处理，提高信息的利用效率和深度。文本自动分类对提高信息搜索的效率和质量具有重要意义：

首先，使信息的分类和运用更加方便快捷。自动对文本进行分类可以为使用者在网络上进行信息检索提供方便。Internet 上的信息资源对于用户来说常常是杂乱且不相关的，用户很难从网络中直接找到所需要的信息资源。为了满足广大用户在信息海洋中方便快捷地获取有价值信息的愿望，这就需要研究有效的文本分类方法，对巨大的网络信息资源进行分类。只有这样，才能提高用户搜索和运用 Internet 信息资源的速度和质量。

其次，可以尽可能减少人力浪费和人为误差的产生。采用自动分类方法可以节省大量人力，而且自动分类的效率是人工分类效率的百倍甚至千倍[2]。由于计算机运算速度快，因此，使用计算机自动对文本进行分类的速度和效率是人工分类所不能及的；而且计算机的计算精度高，减少了人为错误产生的可能性。

目前，对文本分类方法的运用和研究已经成为重要课题，对于文本分类方法的研究已经有很多，但其中还有不少问题值得进一步探讨。例如文本分类系统如何从信息资料中自动学习、获取相关知识；如何在文本分类过程中避免机械的字串匹配，实现接近人脑智能性的语义分类等。针对这些问题所展开的研究，对于提高文本分类的性能，进而

改善人类信息获取的效率，推动社会进步，具有重要的理论意义和现实意义。

本文立足于语言学文献开展研究，主要原因有二：

1．语言学既是一门基础学科，同时又涵盖许多边缘学科，范围之广是其他任何学科所不能及的。

语言学是研究人类语言以及所有人类语言背后规则的科学，按研究目的和范围的不同可以分为普通语言学、个别语言学、历史语言学、描写语言学等。内部又可以分为语音学、语义学、词汇学、语法学、修辞学、方言学等类别。随着科学的发展，语言学不再作为一个独立的学科存在，而是同哲学、历史学、人类学、心理学、逻辑学、文学等密切相关，同数学、声学、数理逻辑、电子学等也建立了紧密的联系。语言学知识不仅对认识语言的本质、特点和发展规律，指导语言教学，确定语言规范，研究翻译理论来说是必要的，而且对了解人类社会发展和民族形成的历史，进行机器翻译，治疗语言障碍的疾病也是有一定帮助的。

随着语言学逐渐成为一个热门的学科，对语言学的研究也逐渐加深。语言学与其他领域交叉产生了一些新兴学科，其中包括社会语言学、心理语言学、认知语言学、应用语言学等多个领域。同时也产生了很多具有语义关联的新概念，例如，用户在查找有关“历时语言学”的文献资料时，基于关键词的文本分类方法只能分析出包含这个词语的文本资料，但是，从语言学专业角度来讲，“演化语言学”和“历时语言学”是同一个概念的两种不同的说法。这时，只有分清楚两词语在语义上的同义关系，才能在检索出有关“历时语言学”文章的同时，也检索出有关“演化语言学”的文章，在很大程度上提高文本分类的查准率和查全率。因此，研究语言学文献的自动分类方法具有重要的理论意义和良好的应用前景。

2．由于作者本人知识水平有限，导致研究工作只能局限于这个领域。但是，真心希望通过作者本人的微薄之力，可以对其他学科的研究提供些许帮助。

1.2 国内外研究现状及分析

1．初级阶段

文本分类可以追溯到上世纪五、六十年代，早期的文本分类主要是基于知识工程（Knowledge Engineering），通过手工定义一些规则对文本进行的分类。应用知识工程

方法在实际操作过程中，最大的缺点和不足就是需要专业人员手工编写分类规则来表达领域专家所拥有的知识，运用这些规则将文档分到一个给定的类别体系中[3]。这种方法不仅需要有领域专家的合作，而且还需要知识工程师手工编制大量的推理规则，具有很大的限制性和不确定性，最能代表这种工作方法的是路透社开发的 Construe 系统[4]。

2．进一步发展

20 世纪 90 年代以来，随着网上在线文本的大量涌现和机器学习的兴起，大规模的文本分类和信息检索再次引起了研究者的兴趣。文本分类系统首先通过在预先分类好的文本集上训练，建立一个判别规则或分类器，从而对未知类别的新样本进行自动归类。它不再需要大量的领域专家的参与，算法也独立于某个领域，不再受到领域知识的限制，能适用于任何领域的学习，使得它成为目前文本分类的主要方法[5]。几种最能代表国外自动分类系统的研究成果如表 1－1 所示：

表 1－1 国外近年来开发的自动分类系统

序号	时间	完成机构	完成人员	技术特点
1	1994 年	At&T 实验室	David D. Lewis 等	基于非确定性的自动分类技术
2	1996 年	At&T 实验室	William W. Cohen 等	电子邮件的自动分类
3	1997 年	德国 Dortmund 大学计算机系	Torsten Joachims 等	基于向量空间模型的自动分类
4	1997 年	美国 Stanford 大学计算机系	Daphne Koller 等	基于很少语料词汇的层次自动分类
5	1998 年	美国 Carnegie mellon 大学计算机系	Yiming Yang 等	采用决策树等聚类算法的在线自动分类
6	1999 年	美国 Just Research 公司	Andrew McCallum 等	运用信息熵理论、Bayes 理论等实现多类号的自动分类
7	1999 年	美国 Massachusetts 大学计算机系	Jamie Callan 等	针对文本库的自动分类系统
8	1999 年	美国 IBM 和 Oracle 公司		为推广电子商务研制基于文本内容的电子邮件自动分类
9	1999 年	Microsoft 公司		为其浏览器开发基于内容属性分类的插件

国内的自动分类研究工作始于 80 年代，经过 20 多年的发展，已经有了一些比较有代表性的辅助归类和自动归类系统。国内比较典型的自动分类系统如表 1－2 所示[5]：

表 1—2 国内近年来开发的自动分类系统

序号	完成时间	完成机构	完成人员	主要技术特点
1	1986 年	上海交通大学计算机系	朱兰娟，王永成	根据原有的类别主题词表和 Bayes 最小损失原则确定分类
2	1995 年	南京大学	苏新宁等	主题词与类号关系表，确定权重系数，分类前控词典，停用词表

目前，对中文文本自动分类而言，主要有三方面的因素影响其分类效果：

1. 虽然国外的英文文本分类方法已经日渐成熟，很多英文文本分类的方法可以借鉴到中文文本分类系统中来，但是，语言方面毕竟存在很大的差异，不能完全照抄照搬。而且随着中文语义，词汇等方面的不断发展，更需要我们开发适用于当前汉语发展的中文文本分类系统。国内外对文本分类的研究大都是围绕对词的统计分析展开的，但是相对于英文来说，中文文本中词语的正确切分是一个很大的难题，分词的正确与否成为影响分类效果的重要因素之一；

2. 另一个影响分类系统正确率的重要因素是词汇差异（Vocabulary Gap），许多文本分类系统采用抽取关键词或类别词的方法对文本进行分类。这样的系统通常都是基于一种假设：类别描述词表与文本之间共享这些词语，我们可以称这种相关性匹配为基于表层的匹配（Surface-Based Matching）[6]。由于几乎不受限制的自由文本用词和受控的类别词表之间存在很大的差异，这种基于表层的匹配不可避免地存在着难以达到更高分类正确率的问题。

3. 文本分类的知识和策略也是影响分类效果的一个重要因素[1]。

1.3 本文组织

本文在传统文本分类技术的基础上，运用本体论的思想，研究语言学文献的自动分类方法。主要包括三个部分的研究内容：

1. 确定语言学文献自动分类的领域，对语言学文献进行预处理；

2．用本体论的思想建立语言学文献的领域本体；

3．将语言学文献的领域本体应用到对语言学文献的分类过程中，力求取得更好的分类效果。

本文共分五章，文章结构及各章主要内容如下：

第1章：引言。介绍文本分类的研究背景和研究意义；分析国内外文本自动分类的研究现状；给出本文的研究工作；最后，介绍本文的组织结构。

第2章：本体论的观点。详细介绍本体的渊源和定义；分析建立本体依据的原则以及本体的组成成分和本体的分类。对本体的概念做出一个全面立体的介绍。

第 3 章：构建语言学文献领域本体所使用的关键技术。首先，采用向量空间模型（VSM）的方法表示文本；其次，利用词或短语之间的概念关联，运用经过一定改进的 TFIDF 算法提取文本特征；最后，介绍了词语之间的几种语义关系。

第4章：语言学文献领域本体的构建。首先确定建立领域本体的范畴和目的；其次，对语言学的相关概念进行处理，确定领域本体的特征项，并采用 Protégé 工具建构语言学领域本体；并对语言学领域本体进行形式化编码；最后，语言学领域本体还要随着社会的发展不断改进和充实。

第5章：基于本体的语言学文献分类过程。首先介绍基于本体的文本分类流程；然后，对语言学文献进行预处理，得到待分类文本的特征项；接下来依赖语言学领域本体对语言学文献进行文本分类，这里使用了基于概念语义和非一致性模糊匹配两种算法，通过评估得出结论：基于本体的文本分类结果准确率高于其他分类方法。由此证实基于本体的文本分类方法切实可行。

第6章：结论和展望。对本文提出的内容进行总结，并提出下一步的工作和目标。

第 2 章　本体介绍

本章首先从理论上介绍本体的渊源、定义、组成、建构本体所依据的原则、当前最流行的本体的分类方法和本体的应用。

2.1 本体的渊源

本体（Ontology）原本是一个哲学概念。17 世纪初，西方哲学家提出“本体”这个概念，用于避免“形而上学（Metaphysics）”中的一些二义性问题；18 世纪初，本体已被哲学界广泛采用。它指的是探究天地万物产生、存在、发展变化的根本原因和根本依据的学说[7]。

近年来，关于本体的研究、开发和应用越来越多。20 世纪 90 年代初期以来，国际计算机界举行了多次关于本体的专题研讨会。并取得一个共识，把现实世界中某个应用领域抽象或概括成一组概念及概念间的关系，构造出一个领域的本体，可以使计算机对该领域的信息处理更为方便，人们在运用这些成果时也更为准确和快捷。本体正逐步成为知识获取以及自然语言处理研究的一个核心内容。

2.2 本体的定义

关于本体的定义，哲学界和计算机界有着很大的差别。在哲学界，本体是表达哲学理论的术语，是指关于存在及其本质和规律的学说，是物质存在的一个系统的解释，这个解释不依赖于任何特定的语言。

而在计算机领域，本体则被解释为一种表达形式。它将领域的知识概念化，并可以表达成计算机能够理解的形式。虽然本体论（或称实体论）这个概念在计算机科学中变得越来越重要，然而，到目前为止，在计算机界却很难为本体论下一个确切的定义。斯坦福大学的 Gruber 给出的定义得到许多同行的认可，即本体论是对概念化的精确描述。本体论的最终目标是精确地表示那些隐含（或不明确的）信息，使得它们可以为计算机领域的发展服务。

2.3 本体的组成

本体研究的是客观事物存在的本质，一个本体就是某个领域或一个领域的某个方面的客观存在的本质。我们可以通过客观存在的概念来认识其本质。首先，客观事物存在于与其相关联的其他事物之间、存在于自身的变化之间；其次，具体的事物与它们之间的关联一起构成具体的存在，对具体的存在进行概括产生抽象的存在；最后，这些客观事物及其之间的关联形成事物的一个概念关系。

本体的组成从形式上说，可以由概念类、关系、函数、公理和实例（属性）5 种元素组成[8]。

1．概念。这里所说的概念是广义的概念，它通常可以构成一个分类层次。概念是客观事物在人脑中的反映，是对事物进行概括的表征。这样的事物可以是抽象的，也可以是具体的。例如，在语言学文献中，“人称代词”就是一个概念，而其中包含的“你”“我”“他（它）”则是这个概念的实例化；

2．关系。关系表示概念之间的一类关联，反映了多个概念之间的内在联系，例如：同义关系是表示两个或两个以上概念之间等同的关系，近义关系则是表示两个或两个以上概念之间相近的关系；

3．函数。函数也是一种特殊的关系，可以用来定义或者计算概念与概念之间、概念与实例之间、实例自身之间的关系；

4．公理。公理用来表示一些永真式，即永远不变的关系或者概念；

5．实例。实例是指属于某概念类的基本元素，即某概念类所指的具体实体，特定领域的所有实例构成领域概念类在该领域内的指称域。

2.4 建立本体依据的原则

从前面的章节中可以看出，这里所说的本体是人为设计的关于某个领域的概念模型的一种表示。Gruber 曾经给出了 5 条设计本体的基本原则[9]。

1．明确性、客观性和完整性：本体应该用自然语言对所定义的术语给出明确的、客观的语义定义，即必须有效地说明所定义术语的意思。而且，当定义可以用逻辑公理表达时，它应该用逻辑公理表示，即形式化表达。 同时，所给出的定义必须是完整的，

能够完全表达所描述术语的含义。Gruber 提出，在可能的条件下，完整的定义（即，同时由必要条件和充分条件表示的谓词）要比一个部分定义（即，仅用必要条件或充分条件定义的谓词）要好。

2．一致性：一个本体应该是前后一致的，也就是说，由它推断出来的概念定义应该与本体中的概念定义一致。由术语得到的推论与术语本身的含义是相容的。至少，所定义的公理以及用自然语言进行说明的文档应该是一致的。

3．可扩展性：一个本体提供一个可共享的词汇，它应该尽可能提供概念的基础，同时，它的表示应该便于人们对这个本体概念进行扩展和进化。

4．编码误差尽可能小：本体应该处于知识的层次，而与特定的符号及编码无关。本体的编码误差应该控制在尽可能小的范围内。

5．最小本体承诺：一个本体应该在提供必须的共享知识的条件下，要求有最小的本体承诺。也就是说，它应该对所模拟的事物产生尽可能少的推断，而让共享者自由地按照他们的需要去运用这个本体，使之专门化、实例化。

除了 Gruber 以外，许多研究者根据自己的实践，进一步提出了其他本体设计原则，如，J. Arpirez 等人提出，本体设计应该遵循以下 3 条设计原则：

1．尽可能使用标准术语；

2．同层次概念之间保持最小的语义距离；

3．可以使用多种概念层次，采用多重继承机制来增加表达能力。

但是，目前还不存在公认的本体设计原则和评价标准以及质量保证标准，所有这些本体设计都是十分笼统和抽象的，因此，这些原则需要我们在实践中根据客观情况的不同，在不一致的原则中间进行权衡，灵活掌握。这也正是我们进行更加深入研究的理由之一。

2.5 本体的分类

目前关于本体的研究日益广泛，尤其是国外。不同的研究机构都建立了各具特色的本体。针对各种不同的本体，也出现了不同的分类方法，主要有以下三种分类方法：

1．根据本体的应用主题分类

根据应用主题的不同，本体可以分为以下 5 类[10]：

（1）领域本体：领域本体在一个特定的领域内可以得到广泛的应用，它提供的是该领域特定的概念定义和概念之间的关系，提供该领域发生的活动以及主要理论和基本原理等。对特定领域的本体研究和开发目前已经涉及许多领域，包括企业本体、医学概念本体、酶催化生物学本体、陶瓷材料机械属性本体。

（2）知识表示本体：研究重点是语言对知识的表达能力。典型的有斯坦福大学知识系统实验室提供的一种称为知识交换格式（KIF, Knowledge Interchange Format）的知识描述语言，以及可以在线将各种知识转换为 KIF 的本体服务器 Ontolingua。目前普遍认为，所有其他的知识表示形式都可以转换为 KIF 的形式。

（3）通用和常识本体：关注于常识知识的使用。中国科学院数学所承担的国家自然科学基金重点项目“常识知识的实用研究”中开发的结合 Agent 和本体的知识库 Pangu 也属于通用知识本体的研究范畴。

（4）任务本体：也称为方法本体，是本体研究的另一个分支，主要研究可共享的问题求解方法，这里的推理方法与领域无关，任务本体主要涉及动态知识，而不是静态知识。具体的研究主题包括：通用任务、与任务相关的体系结构、任务方法结构、推理结构和任务结构等。

（5）语言学本体：是指关于语言、词汇等的本体。典型的实例有 GUM(Generalized Upper Model)和普林斯顿大学研制的 WordNet。

2．根据本体表示的形式化程度分类

根据表示的形式化程度不同，本体可以分为以下 4 类[11]：

（1）完全非形式化：完全采用自然语言表示，结构非常松散，典型的有术语列表。

（2）结构非形式化：采用受限的或结构化的自然语言进行表示，能有效提高本体的清晰度，减少二义性。如，Enterprise Ontology 的文本版本。

（3）半形式化：采用一种人工定义的形式化语言进行表示，目前已有许多研究机构开发指定了这类形式化本体表示语言，采用 Ontolingua 描述的本体都属于这一类。

（4）完全形式化：所有术语都具有形式化的语义，并能在某种程度上证明包括一致性和完整性方面的属性。

3．根据研究的层次分类

本体的研究和开发工作是在不同的层次上进行的。根据本体的研究层次，可分为[12]：

（1）顶层本体：主要研究非常通用的概念，如空间、时间、事务、对象、事件、行为等，他们完全独立于特定的问题或领域。因此可以说顶层本体是在一个很大范围内的知识层次。

（2）领域本体：研究与一个特定领域相关的术语或关系。

（3）任务本体：定义通用任务或推理活动。任务本体和领域本体处于同一个研究和开发层次。它们都可以应用顶层本体中定义的词汇来描述自己的词汇。

（4）应用本体：描述特定的应用，它既可以应用特定的领域本体中的概念，又可以引用出现在任务本体中的概念。

2.6 本体的应用

本体构建的目的就是应用。这方面的研究遍布于文本分类、人工智能、信息管理、知识管理相关的各个领域，典型的应用有：

1．基于语义的文本分类和信息检索，特别是网络搜索引擎和数字化图书馆。在信息检索领域和数字化图书馆中，加入本体的思想，可以在检索过程中更加准确的对文本进行定义和分类，快速找到相关的信息。例如，在信息检索过程中，输入检索词语“第一语言教学”，加入本体论的思想后，有关“母语教学”的文章也会出现在检索结果中，可以在很大程度上提高网络信息的利用率。

2．基于本体的数据集成、机器学习等。数据集成和机器学习需要了解某个领域的全部知识，本体思想的引入，可以对数据集成和机器学习提供一定的便利。

3．领域本体的应用。在各个不同的领域建立不同的本体，这样就可以有针对性分析事件，提高研究的效率。

4．语义 web 服务。语义 Web 是 Web 未来的发展趋势，本体技术提供了语义 Web 描述词汇的精确定义，为真正实现 Web 信息的语义表示奠定了基础。

5．在线元数据管理和自动信息发布。在线元数据管理和自动信息发布是一个实时的信息处理过程，有了本体的参与，可以使准确率得到进一步提高。

2.7 本章小结

本章主要介绍了有关本体的理论知识，了解了本体的概念来源于哲学，随着科学的

发展正在被广泛应用于科学研究的各个领域；介绍了研究者对本体概念的不同理解，目前比较认同的观点是本体论是对概念化的精确描述；本体的组成包括概念类、关系、函数、公理和实例 5 种元素；还分析了建立本体必须依据的原则，即明确性、客观性、完整性、一致性、可扩展性、编码误差尽可能小、最小本体承诺等，为下一章中语言学领域本体的建构奠定了理论基础。另外还介绍了在三种不同的分类标准下对本体的分类情况；最后简单地说明了本体的应用情况，使我们对本体的概念产生了一个全面立体的认识。

第 3 章　构造领域本体所使用的关键技术

基于本体的文本分类能否顺利实现，主要取决于领域本体的构建。而领域本体构建是否成功，又是由其中若干个关键技术的选择和运用所决定的。主要包括：文本表示、特征项粒度选择、特征提取和语义推理。

3.1 文本表示

计算机不能识别人类的语言，所以需要文本表示的过程，把人类的自然语言变成计算机可以看懂的符号。目前，在信息处理过程中，文本的表示大多数采用向量空间模型（Vector Space Model, VSM）的方式[13]。

本文采用向量空间模型的方式来表示文本：给定一个自然语言文档 D，在选定了特征项以后，用 $D=(s_1,w_1; s_2,w_2; \cdots; s_N,w_N)$ 来表示文档 D，其中 $s_i=(i=1,\ldots, N)$ 为特征项，w_i 为 s_i 的权重，规定 $s_i=(i=1,\ldots, N)$ 互不相同。把向量 $D(w_1,w_2,\ldots, w_N)$ 叫做文档 D 的向量表示或者向量空间模型，文本用向量 D 来表示。

接下来，要对文本进行分词处理。中文分词一般采用最大匹配法[14]。最大匹配法是机械分词方法的一种，按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配，如果能够在词典中找到某个字符串，则匹配成功，识别出这个词语。按照由左到右的方向匹配的方法叫做正向最大匹配法，由右到左的方向进行匹配的方法叫做逆向最大匹配法。双向最大匹配法（正向最大匹配法和逆向最大匹配法的结合）速度快、词表开放、格式简单容易扩充。

对分词结果进行词性标注，这个过程中还要完成对文本的去虚词处理，建立一个有序的虚词词表是前提。对从分词结果中提取到虚词采用二分查找来判断这个词是否在虚词词表中，如果在则丢弃；如果不在则保留。

3.2 特征项粒度选择

使用向量空间模型方法表示文本，并对文本分词和词性标注以后，下面就要对文本进行特征抽取。采用什么标准抽取特征项，对领域特征项的确定具有很大的影响。一般的特征项的抽取有三种粒度可供选择：一个是字，一个是词或短语，一个是概念特征。

1．字。使用字特征的特征抽取过程最简单。那是因为在国家标准 GB2312－80 中定义的常用汉字仅为 6763 个，由于这些常用汉字数目比较少，所以抽取过程所需的时间和空间的开支都不会很大，效率也比较高。但是，就字本身而言，对文本表示的功能性很差，根本无法独立完成对文本特征的表示，更不能准确的表达文本之间的语义信息，不可取。

2．词或短语。词是汉语中能够准确表达语义信息的最小的语言单位。与字相比，在语义表达方面的优势显而易见。如果不计算专业领域的词汇，通用的词和短语有 10 万左右，使用词特征就要面临复杂的分词问题，而同时，并不是说所有的词语和短语都适合作为特征项。在词频统计时，会出现很多频率很高，但是对文本特征表现力却很弱的词语或短语，就不能作为特征项；相反的，也会有很多出现频率不高，但是却可以一词定类的词或短语，比如“复元音韵母”，只要出现这个词语，几乎就可以判定文本的特征，从而确定文本的类别。由此可见，使用词或短语的特征进行特征抽取具有很大的不准确性。

3．概念特征：词语之间往往都存在同义关系、从属关系、近义关系等丰富的语言现象。理清这些概念层次之间的关系，综合以上词和短语的优点，就可以对文本特征进行很好的抽取，是一种比较科学的特征抽取方法。

综合比较以上三种特征抽取粒度的优劣，本文采用概念特征的标准对文本特征项进行选择。这样既可以避免字特征文本表示功能较弱的缺点，又可以在一定程度上弥补单纯靠词或短语表示文本特征的不确定性，从而保证抽取特征项的准确性和科学性。

3.3 特征选择

经过分词以后的文本，词汇量很大，而且用 VSM 表示的文本，向量空间的维数很高，不利于计算机处理；另外，每个词语对表现文本主题的贡献程度不一样，有些词出

现频率很高，但是对确定文本类别没有太大帮助；有些词出现频率虽然不高，但是却可以一词定类。这就需要一个特征选择的过程。

特征选择的基本思想是：在对文本中的关键词进行选择之后，计算每个词语的词频，并将经过学习预设的重要度作为权值，对所有的特征按照其权值的大小排列，通过设定阈值或限定维数，可以得到文档的特征集。由此可见，特征选择是建立在词频统计和计算权重的基础上的。经过词频统计和权重计算，就能生成文本类别的核心向量，这些向量中的特征词可以认为是能代表该类文本特征的类别领域词 [15]。

TFIDF 方法是文本分类过程中特征提取使用最多的方法之一。其中，TF：Term Frequency 为频率因子，表明文档中出现该特征项的频度；IDF：Inverse Document Frequency 为特征项倒排文档频率，表明特征项在文档集合中分布情况的量化。一般的特征选择过程中都是采用这种方法来构造词语权值评价函数，在系统中采用的是由人工分类好的训练语料，让系统进行分析提取。为了保证语言学文献中的常用词（例如“形式名词”中的“名词”词条）得到选择，降低 IDF 的影响，同时为了得到一个单位空间向量，还要对特征向量的各个参量进行归一化处理。

本文采用改进的 TFIDF 算法对文献进行特征选择。首先本文以概念特征作为特征选择的标准，那么给定两个词语，计算它们之间的语义距离。这里，把语义距离定义为两个词对应的属性或概念在特征库中的最短距离。如果两个词中有一个词的属性无法在特征库中找到，或者两个词的属性分别处于两个不同的特征库，就可以认为这两个词之间的语义距离为∞[16]。

设两个词U、V之间的语义距离为p，那么U、V之间的相似度可以用公式(1)来计算：

$$s(U,V)=\begin{cases}H-(p\times(H-L)/D) & (p\neq\infty)\\ 0 & (p=\infty)\end{cases} \tag{1}$$

这里的H和L是两个词之间相似度可能取得的最大值和最小值。在这里，令H=1，L =0。D是U、V所在的特征库中两个实例的语义距离可能的最大取值。即如果某个特征库中深度最大的两个实例或属性的深度分别为D_1、D_2，那么这个特征库的D $=D_1+D_2$。注意，根据上面所说，当$p\neq\infty$时，U、V的实例或属性必定是在同一特征库中，

因此，关于 D 的定义是合理的。以此类推，就可以得到包含实例或属性之间具有语义关系的特征库。

3.4 语义推理

基于本体的分类过程，必须以特征项之间的语义关系作为基础，如何确定词语之间的语义关系呢？这里就涉及到一个概念：语义推理。

语义推理就是通过扩展词语之间的语义关系来确定领域特征项，将所有隐含的信息都显式地描述出来，以此来构建领域本体库。语义关系是建构本体特征库过程中，联系概念与实例的中心环节，因此作为特征库中的联系各级节点之间的纽带而存在。这样的语义推理完成了对元数据概念的语义扩展，主要包括：

（1）同义词关系（Synonym）扩展：同义词是意思相同或非常相近的两个或多个词语，它们之间往往可以相互替换。如“声调”和“音调”、“复元音韵母”和“复合元音韵母”等。

（2）上下位关系（Hypernymy/Hyponymy）扩展：就是包含与被包含的关系。下位词是上位词的特例，如“声调”和“阴平、阳平、上声、去声”之间的关系，其中“声调”是上位词，“阴平、阳平、上声、去声”是下位词。在分类过程中，有时通过概念的上下位概念也能分析出潜在的有用信息。

（3）相似词扩展：相似的两个词之间具有兄弟关系，但不是同义词或者上下位词，如“阴平”、“阳平”、“上声”、“去声”四个概念相互之间的关系。

（4）歧义概念的标注：自然语言中存在很多一词多义的现象。为了排除歧义的干扰，我们借助文档特征进行唯一标注，这样就可以达到消除歧义的效果。

这些初始概念经过语义分析，可以防止概念的冗余，避免重复的概念，并且通过领域专家的确认后，可以成为领域特征项，作为本体的核心概念或者实例，在本体库建构过程中确定下来，在以后的不断完善过程中还可以作为新的特征项，源源不断地扩充进来。

3.5 本章小结

本章分析了建构领域本体所需的关键技术，在用向量空间模型对表示文本的前提下，从概念特征出发对文本进行特征选择。采用经过一定改进的 TDIDF 算法，确定领域特征项。还应该明确概念之间的语义关系，包括上下位关系、同义关系、近义关系等。只有首先明确构造领域本体所需要的关键技术，才能为领域本体构建工作提供技术支持，有利于研究工作的开展。

第 4 章　语言学领域本体的构造

4.1 构造领域本体的必备条件

4.1.1 本体形式化描述语言的选择

本体形式化描述语言直接影响本体模型的表达能力和可扩展能力。目前的形式化本体描述语言非常多，主要有 RDF 和 RDF-S、OIL、DAML、OWL、KIF、SHOE、XOL、OCML、Ontolingua、CycL、Loom[17]。经过比较，我们选用了 OWL（Web Ontology Language）。

OWL 的优点是以 Web 资源为描述对象，具有良好的应用前景。另外，OWL 是基于描述逻辑的，所谓描述逻辑（Description Logic, DL）是一阶谓词逻辑的可判定子集，能够提供可判定的推理服务，并且具有语义特征[18]。这就意味着基于描述逻辑的 OWL 的函数和公理都有相应的逻辑描述表示，利用 OWL 构建的本体库除了具备良好的表现能力外，还具有强大的推理能力。这对于 Web 资源的逻辑检测、本体集成、知识整合是非常重要的。

4.1.2 本体开发工具的选择（Protégé+OWL plugin）

目前国内外已经有许多成熟的本体开发平台软件可供选择。经过我们对部分常见工具的试用与比较，选择的是其中的佼佼者 Protégé3.2.1[19]（用户界面截图如图 4-1 所示）。Protégé 是由斯坦福大学医学信息化研究小组开发的，一个基于 Java 环境、开放式架构的开源知识建模工具[20]。其扩展的 OWL 插件是目前最为强大的 OWL 本体构建工具。Protégé 不仅具有良好的可扩展性和简单灵活的用户定制界面，还具有如下一些特性：

1. 支持图形化本体编辑模式；

2. 支持数据库存储模式；

3. 基于 OWL 数据库的多人开发模式和支持逻辑检测功能等。

最新版本的 Protégé 还增加了对资源多语言描述的支持。更为重要的是，Protégé 还拥有超过 50000 人的注册用户和邮件列表用户，高效的技术服务支持以及丰富的技术

资料和本体资源。这些都极大地方便了我们本体构建的学习和问题的解决。

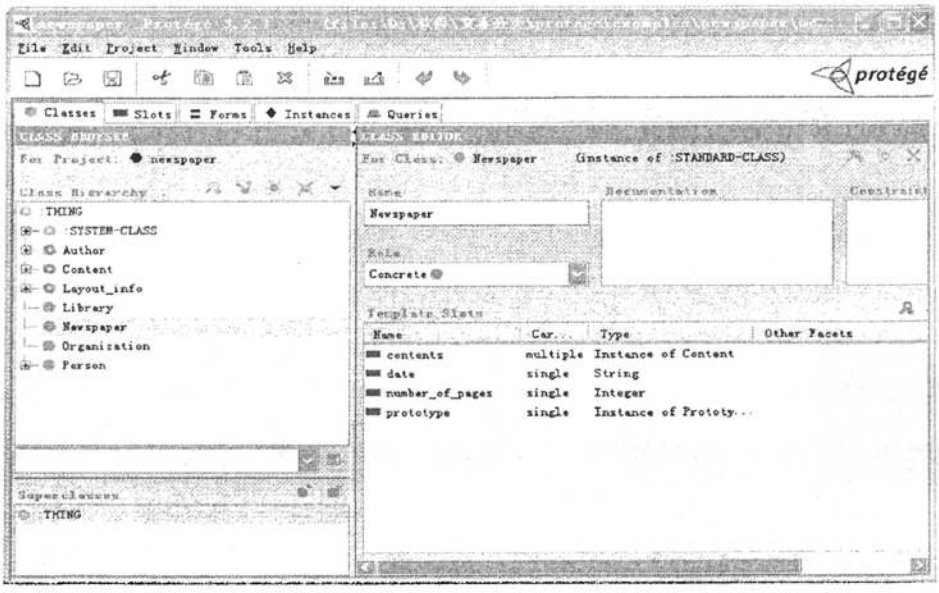

图 4—1　Protégé3.2.1 用户界面截图

4.1.3 确立本体建构的指导思想

本体构建是一个复杂的系统工程，需要明确的开发思想、方法和规划来推动项目。近年来，借鉴软件工程的思想和方法已经被大家广泛接受。不少国内外专家提出了本体构建的方法论：如 IDEF-5、Skeletal Methodology（骨架法）、企业建模法（TOVE）、METHONTOLOGY 以及 Cyclic Acquisition Process 方法等[21]。

4.1.4 领域专家的参与

领域本体构建是本体开发人员与领域专家共同努力的结果。开发人员虽然具有丰富的本体知识和较强的开发能力，但是对特定领域知识却知之甚少，很难建立起面向特定领域的本体模型。所以本体构建非常需要领域专家的参与。

4.2 语言学领域本体的构建过程

从某种意义上说，构建领域本体是一类新的软件活动[22] [23]。领域本体是具体领域

中的概念和关系的抽象描述，是整个语言学文献分类系统的基础，领域本体的好坏程度，直接影响文本分类系统的功能。在此，我们借鉴软件工程的思想和经验，运用基于本体生命周期的构造方法来建构语言学领域本体。

把领域本体的构造过程分为五个阶段，即确定领域本体的目的和范畴、领域本体分析、领域本体的编码和表示、领域本体的进化和扩展以及评测[24]。如图 4—2 所示。

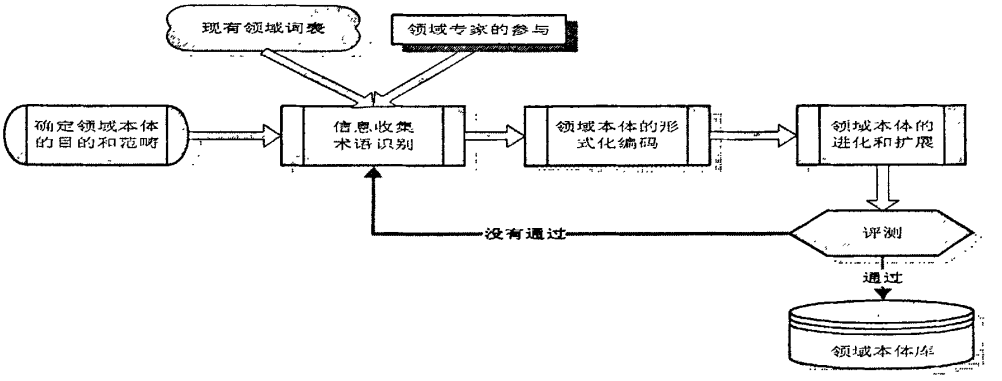

图 4—2 领域本体建构流程图

4.2.1 确定领域本体的范畴和目的

1．建立语言学的领域本体，必须采用语言学文献资料作为语料进行研究。本文选择《现代语文》（语言研究版）作为研究语料。该期刊将文献分为“语文百家访谈”、“语言理论研究”、“语言应用研究”、“语言教学研究”、“语言新观察”以及“咬文嚼字”、“工具书补正”、“词典评析”、“语林考古”等几个部分，大致上涵盖了语言研究的各个领域，同时包含了大量语言学领域比较有代表性的概念，为本文研究工作的开展提供了很大的便利。

本文从中国学术期刊网下载了《现代语文》（语言研究版）2006 年全年的文章共计 909 篇作为研究语料。由于期刊文献涉及范围十分广泛，既包括普通语言学领域，也包含了很多语言学相关领域。在整个语言学领域建立本体的工作过于繁杂，加上作者本人知识水平有限，因此，在这里经过细致筛选，最终确定 150 篇普通语言学文献作为开展研究的对象。选择其中的 100 篇（注意：这里的 100 篇文献必须涵盖普通语言学的各个分支学科）作为取材范畴，其余的 50 篇在后面的分类实验中作为待分类文本使用。

2．明确建立本体的目的。建立本体的目的是：

（1）利用本体思想和 OWL 语言组织和描述语言学领域知识；

（2）建立一个具有逻辑检测和可扩展性的领域本体库，可以更明确、清晰的认识领域特征项之间的语义关系；

（3）实现应用中的快捷分类以及查询，便于使用者在运用过程中更加方便快捷的找到自己所需要的内容。

3．明确目标用户。主要有两类：一类是语言学文献的使用者，一类是对语言学文献进行文本分类的研究人员。

4．最终我们建构的语言学领域本体应该具有以下特征：

（1）简单、良好地定义概念的层次结构；

（2）取得结构可重用性和语义关系表现力之间的平衡；

（3）具有建立在概念公理基础上资源的可扩展性；

（4）支持互操作和多语言特性。

4.2.2 确定核心概念的方法

确定本体构建范畴和目的后，接下来要做的是利用本体建立领域知识概念模型。目前建立领域本体概念模型通常有三种方法[20]：

1．自顶向下（top-down）方法，其表现形式是由现有的领域本体模型构建应用本体模型，其中应用本体为针对特定对象而生成的本体；

2．自底向上（bottom-up）方法，其表现形式为将领域知识中名词性的概念、术语等进行识别、处理二义性、归纳、聚类、泛化等处理，建立概念模型；

3．核心扩展（middle-out）方法， 其表现形式为由具有本体雏形的一组核心概念入手，不断扩展本体概念模型。

其中第 1 种和第 3 种方法在目前的本体构建项目中应用比较多，第 2 种方法适用于拥有大量领域知识资料并且能够使用自动或者半自动本体采集生成工具的情况。采取自底向上方法过于繁琐，而且目前的自动或者半自动本体采集生成工具使用效果也不好。综合上述三种方法的优缺点，本文决定采用核心扩展的方法确定核心概念集。

4.2.3 语言学领域本体特征项分析

首先，利用已有的分词系统对作为源语料的 100 篇普通语言学文献进行分词处理，得到一个含有权重标记的关键词排序；利用第 3 章 3.3 介绍过的特征选择方法，对关键词进行特征提取；并进一步在领域专家帮助下，利用核心扩展的方法产生所有潜在的核心概念。核心概念必须满足没有二义性、互不相交和并集覆盖整个普通语言学领域知识的要求。经过识别、分析和统计得出：

1．首先确定了“语音学”、“词汇学”、“语法学”、“方言学”四个核心概念。作为本体模型的顶级概念，

2．由于本文只是研究传统意义上的语言学概念，因此，这里只是对各个领域做定性研究。这样，在语音学领域，包括“音素”、“音节”等核心概念；词汇学领域，包括“语素”、“词”、“短语”等核心概念；语法学领域包括“性”、“数”、“格”、“时”、“体”、“态”、“人称”、“词类”等核心概念；方言学领域包括“北方方言”、“吴方言”、“闽方言”、“赣方言”、“粤方言”、“客家方言”、“湘方言”等。综上所述，在顶级概念之下，产生了 20 余个二级概念。而这些概念与顶级概念之间都是包含关系，并且满足互不交叉，没有二义性。这样就可以去确定这些概念作为领域本体的核心概念。

3．在这 20 余个二级概念下，又包含着很多三级概念。如：“元音”、“辅音”等作为语音学领域中“音素”的三级概念存在；“单数”、“双数”、“多数”等作为语法学领域中“数”这一核心概念的三级概念存在。经过统计分析，可以得出，类似这样的三级概念在普通语言学领域有 60 多个。

4.……

5.……

……

通过对 100 篇源语料进行分析，对不同级别的核心概念进行统计可以得到普通语言学中存在近 500 个核心概念。本文对概念的统计分析只是作为构建语言学领域本体的前期工作，这里不再做过多的描述。

4.2.4 语言学领域本体层次描述

第一层次：“语言学”是本课题研究的最大的领域，在领域本体库中作为父节点存在。如图 4－3 所示：

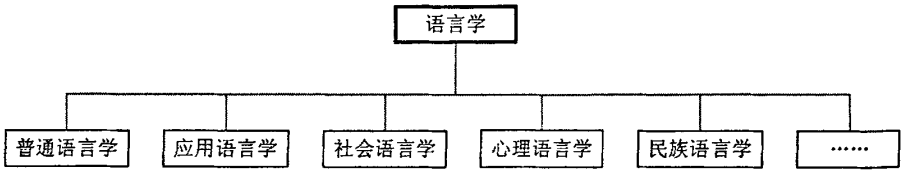

图 4－3 语言学文献领域本体第一层次

在这一层次中，“普通语言学”、“应用语言学”、“社会语言学”、“心理语言学”、“民族语言学”等作为领域本体库中“语言学”这一父节点的一级子节点存在。

第二层次：“普通语言学”、“应用语言学”、“社会语言学”、“心理语言学”、“民族语言学”等分别作为一级父节点存在。如组图 4－4 所示

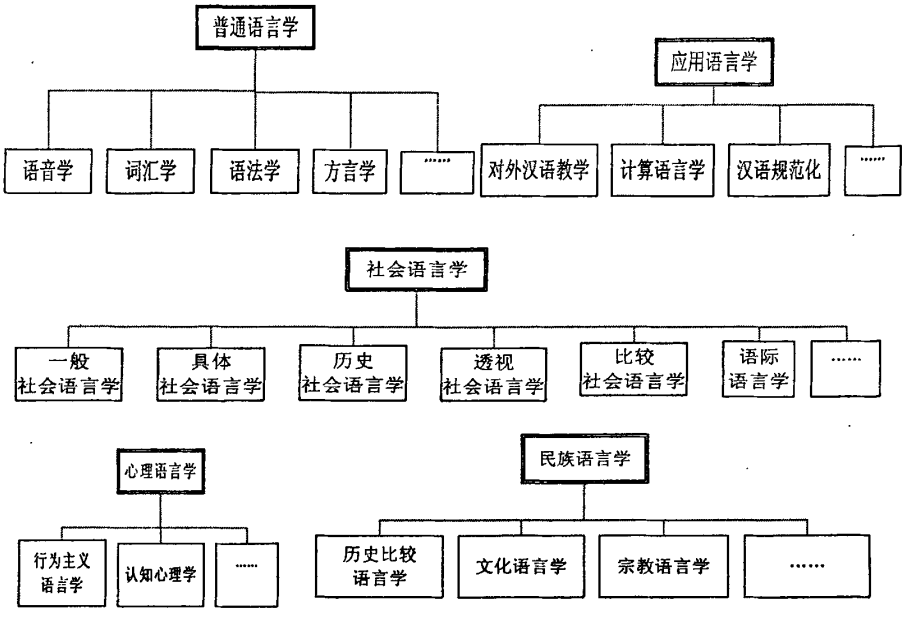

图 4－4 语言学文献领域本体第二层次

“语音学”、“词汇学”、“语法学”、“对外汉语教学”、“汉语规范化研究”、“计算语言学”、“一般社会语言学”、“具体社会语言学”、“历史社会语言学”、“透视社会语言学”、“语际语言学”、“行为主义语言学”、“认知心理学”、“历史比较语言学”、“文化语言学”、

“宗教语言学”等都是作为上一级父节点的子节点存在。在这里值得一提的是，随着“方言学”研究的不断深入，很多学者把“方言学”看成是普通语言学的一个分支学科，使之与“语音学”、“词汇学”、“语法学”并列作为普通语言学的重要组成部分。本文正是采用了这样一种学术界比较认同的观点，把“方言学”提出来作为普通语言学的分支存在。

第三层次：“语音学”、“词汇学”、“语法学”、“方言学”、“对外汉语教学”、“汉语规范化研究”、“计算语言学”、“一般社会语言学”等上一级子节点在这一层次中作为父节点存在。如组图 4－5 所示：

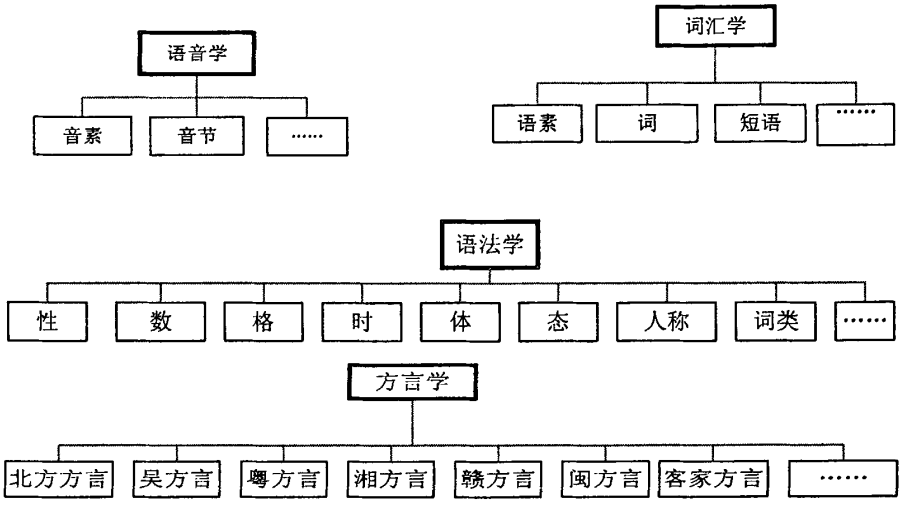

图 4－5 语言学文献领域本体第三层次

由于篇幅有限，这里不能一一列举领域本体这一层次的所有子节点，通过观察上图一目了然，“音素”、“音节”、“语素”、“词”、“性”、“北方方言”等，都是分别作为“语音学”、“词汇学”、“语法学”、“方言学”的子节点存在。

第四层次：“音素”、“音节”、“词”、“短语”等上一层次中作为子节点存在的特征项，在这里作为父节点存在。下面只对“音节”和“短语”所包含的子节点进行说明，如图 4－6 所示：

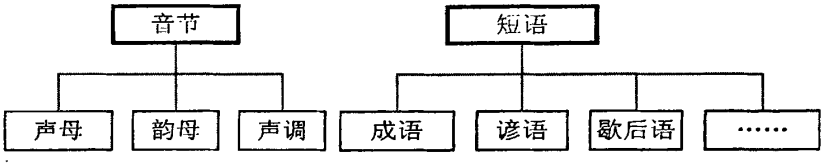

图 4－6 语言学文献领域本体第四层次

在这里，“声母”、“韵母”、“成语”等作为“音节”、“短语”这一父节点的子节点存在。

第五层次：“声母”、“韵母”、“声调”等上一层次的子节点，在这里作为父节点存在。下面只对“声调”所包含的子节点进行说明，如图 4—7 所示：

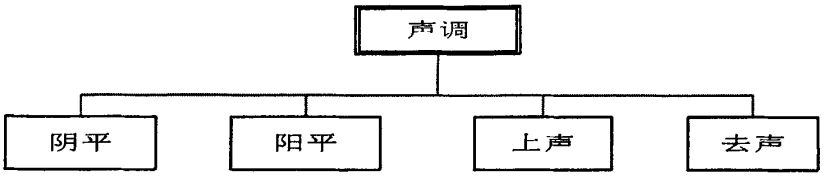

图 4—7 语言学文献领域本体第五层次

在这一层次中，“阴平”、“阳平”、“上声”、“去声”作为“声调”的子节点存在，同时也是这棵领域本体特征树的最后一个概念环节。虽然在这里，上述四个实例不能再进行分解，但是，其他概念还包含很多不同的层次和实例，由于篇幅所限不能一一列举。

第六层次：……

第七层次：……

……

由于本文工作的重点是把这些概念通过一定的语义关系，构建一个语言学领域的本体，下面，对领域本体的构建过程做详细描述。另外，由于篇幅有限，不允许把整个普通语言学领域的概念关系完全描述出来，这里只截取了语音学领域本体的结构图，对语言学领域本体进行描述。

4.2.5 建立概念层次结构

领域本体种的概念层次主要有三部分构成：概念、关系和实例。这里对语音学方面的概念关系进行简单的说明。

1．概念：概念是对客观事物本质的描述，是对客观事物的反映。例如：“语音学”是一个概念，是研究普通语言学语音方面知识的一个类别，作为一个大的聚类存在。

2．关系：在上一章已经详细介绍过，关系说明的是概念之间的联系。两个或两个以上的概念以何种形式联系在一起。例如：“声调”和“音调”之间是同义关系；“元音”和“辅音”之间是兄弟关系，或称为相似关系；“韵母”和“复元音韵母”之间是上下

位关系。

3．实例：用来说明概念的实体。例如：“音素”和“音节”，作为两个实例在说明“语音学”这个概念。而“阴平、阳平、上声、去声”四个实例则是在说明“声调”这个概念。

当然，概念与实例，概念与概念，实例与实例之间的语义关系也可以有很多层次，正是这样复杂的联系构成了语言学领域本体库。

领域特征项的概念类表示为父节点，实例表示为子节点，依靠概念之间的语义关系来连接。本文使用 Protégé 系统对语言学领域本体库进行构建，如图 4－8 所示。

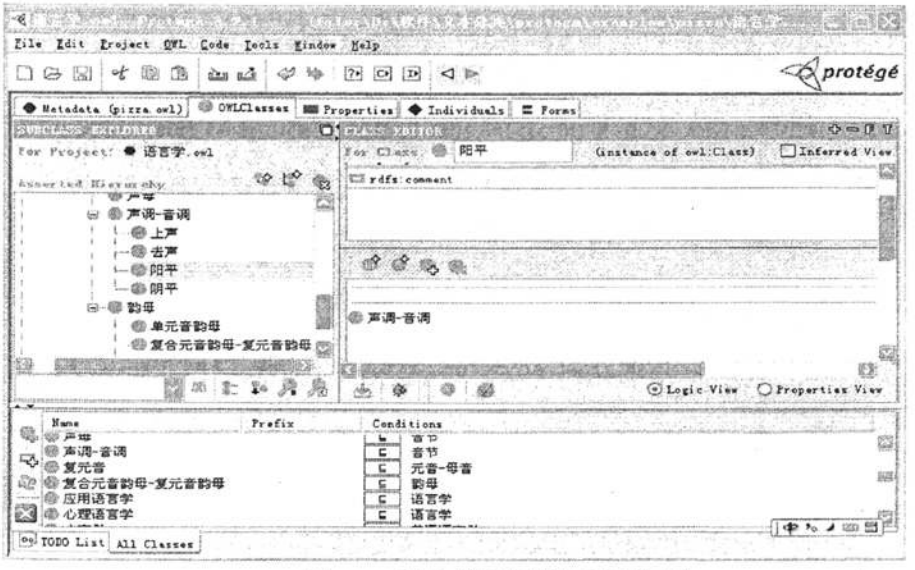

图 4－8 语言学领域本体结构图

由图 4-8 可以看出：

1．“语音学”：在这个本体片断中以父节点的形式存在；

2．“音素”、“音节”作为“语音学”的一级子节点存在。与“语音学”的关系是上下位关系；

3．“元音”、“辅音”、“声母”、“韵母”、“声调”作为下一级子节点存在。与上一级关系“音素”、“音节”的关系是上下位关系；

4．“母音”、“复合元音”、“子音”、“复元音韵母”、“音调”作为原领域特征项的同义扩展项存在，它们与其同义词语之间的语义距离为 0。关系：同义关系；

5．在同一个平面上的概念之间的关系是相似关系。如“应用语言学”和“心理语

言学”等。

依此类推，就可以在语言学领域建立一个支持整个领域知识的本体库。这个过程中，机器只是起到了辅助作用，人工才是根本。通过领域专家的分析和确认才能确定词语之间的语义关系，添加到领域本体中来。

4.2.6 形式化编码和表示

利用 Protégé 系统建立领域本体以后，领域特征以.owl 为后缀的形式保存下来。图 4－9 是语言学领域本体的一段编码。“语音学”在这里作为一个大的聚类存在，它统率所有小的概念，是父节点，也是相对来说比较大的范畴。“音素”、“音节”作为它的子节点存在，同样以.owl 为后缀保存在领域本体中。由于领域本体的形式编码属于系统自带的功能，这里不再做过多解释。

```
<owl: Class rdf:ID="音素">
    <rdfs:subClassof rdf:resource="#语音学">
</owl:Class>
<owl: Class rdf:ID="音节">
    <rdfs:subClassof rdf:resource="#语音学">
</owl:Class>
```

图 4－9 领域本体的形式化编码

4.2.7 语言学领域本体的进化和扩展

本文采用机器计数和人工判断相结合的方式，把新概念加入到本体库中，实现本体扩展。如果使用者输入的语言学文献的关键词不在当前领域本体库中，但是在文本分类过程中被选定为文本特征项的次数很多，又是能够代表领域特征的词语，那么，计算机会提示系统管理人员这个词可能是一个新的概念，这时候由系统管理人员对这个词进行分析，有必要的话，还要请领域专家进行鉴别。如果可以看作是一个新的概念，则加入到本体库中适当的位置；如果不能看作是一个新的概念，则舍弃。

4.2.8 语言学领域本体的评估

本文参考 Mariano 评估和比较本体性能的模型标准[25]，制定以下评估内容对领域本

体库的建立进行有效的评估。

1．本系统所建构的领域本体是与应用密切相关的，正是因为语言学文献的使用者在对所需语言学文献的查找过程中感到一定的困难，现有的文本信息、专家知识、通用本体和文本分类工具不能很好的满足语言学文献使用者的需求。本系统是在这种需求达到一定程度的基础上构建的。因此，可以说是应时而生的。

2．本系统中对信息的收集，主要是在有代表性的期刊论文中进行的。选择了普通语言学方面的 100 篇文章作为源语料，可以比较全面的代表整个普通语言学的语义特征，具有较强的表征性和说服力。

3．对语言学文献特征的提取是采用比较精确的经过一定改进的 TFIDF 算法计算获得的。再加上领域专家对特征项语义的扩展，使领域特征项的选择在准确性方面有了一定的保障。

4．在形式化编码的过程中，本文根据语言学词汇语义之间的关系进行编码，采用 OWL 语言存储到数据库中。对语言学文献本体库的稳定性和准确性能够发挥一定的保障作用。

在领域的本体的评测阶段，一般采用领域专家的观点对特征项进行分析，如果能通过专家的评测，则添加到领域本体库中；如果不能得到领域专家的认可，就要回到信息收集和术语识别的过程，重新选择特征项。

4.3 本章小结

这一章介绍了建构语言学领域本体的过程。首先，对语言学文本进行特征选择，抽取其中最能代表领域特征的项，作为领域本体的核心，利用 Protégé 系统，在细致分析语言学文献特征的基础上，建立语言学领域的本体库，为下一步文本分类工作提供了必要的基础条件。另外，因为语言是一个开放的系统，可以根据社会的发展不断的扩充修订，那么，语言学领域的本体库也随之成为一个开放的系统，可以随着领域特征项的变化而变化，无论是领域概念、实例还是概念和实例之间、概念和概念之间、实例和实例之间的关系都会不断改变。比如，随着社会的不断发展，语言学与地理学科结合，产生了地理语言学，随之也会产生很多该领域的特征项，“地理语言学”就可以作为一个概念扩充到领域本体库中来，而它包含的很多领域特征项，也可以作为实例添加进来。

最后，还谈到了领域本体的评测，随着语言学研究的不断深入，语言学领域本体库必将为文本分类和信息检索提供很大的便利。

第 5 章　基于本体的语言学文献分类过程

利用传统的分类方法对语言学文献进行分类，用户要么必须了解整篇文章的知识背景，要么必须了解语言学相应的关联语法。而这样的专业知识，对一般用户来讲，具有很大的领域局限性。除此之外，还存在两个用户本身无法克服的困难：

一、 忠实表达。所谓“忠实表达”就是有时用户不能很清楚的表达自己想要分类的内容，只了解所要分类内容的一个大概印象，比如，当需要查询“语言学分支”时，用户自己心里明白想要查询具体的分支内容，即：语言学分支包括社会语言学、心理语言学、应用语言学等。但是，在传统的分类系统下，用户输入“语言学分支”这个关键词却得不到想要的答案，而多是一些跟“分支”关系不大的结果；

二、 表达差异。所谓“表达差异”指人类的自然语言中，随着时间、地域或领域的改变，同一个概念可以用不同的关键词来查询（一个典型的例子，在用户想要获得“隐喻”的概念时，与其同义的“暗喻”则不能出现在结果中）或者一个关键词可以表达多种概念（这种问题虽然在语言学文献的分类过程中很少出现，但是，就系统的通用性而言，这不能不说一个很大的缺陷）[26]。

随着社会的发展，传统的文本分类方法，在这些因素的影响下，会更加凸现其弊端。因此本文在原有分类方法的基础上，提出了一种基于本体的语言学文献分类方法，充分综合领域知识优点，使用领域本体中概念之间的语义关系和层次关系进行综合匹配，明确文本所属类别，从而部分地克服上述问题，得到较好的分类结果。

5.1 分类流程设计

本文所设计的分类流程，是建立在语言学文献领域本体库基础上的。如图 5－1 所示。该系统由 2 个层次组成：本体层和分类层。本体层是上面的一层，在本体层中，领域本体提供了相关领域的知识，它为文本的特征提取提供了保障[27]。分类层是下面的一层，则是对一个未知文档使用分类工具进行分类的过程，其中，正是因为有了领域本体的参与，才有可能使分类系统更加合理化、科学化。

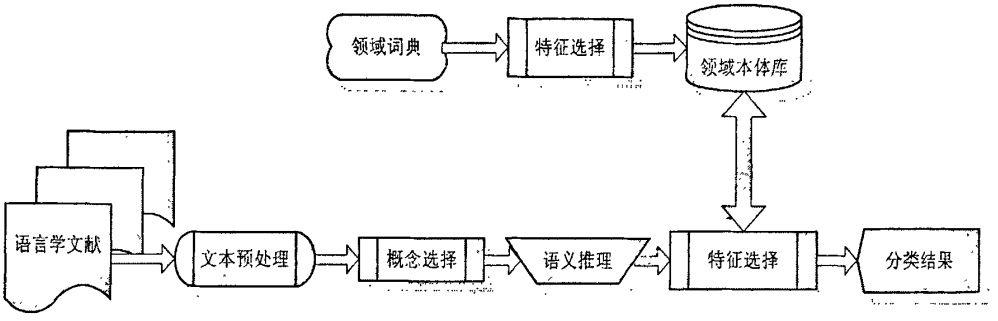

图5－1 基于本体的语言学文本分类流程图

5.2 分类方法选择

文本分类的具体实施过程有多种方法，如图 5－2 所示。主要可以分为两大类[28]：参数算法[29]和非参数算法[30]。参数方法也可以称为基于外延的分类方法，它是假定一个文本的概率分布模型，通过训练得到具体参数的估计值，以此数值为依据来判定文本所属类别。非参数方法也可以叫做基于内涵的分类方法，它不假定任何概率分布模型，通过准则函数直接对文本进行处理，得到各类的权重向量，然后对待分类样本进行分类。由于分类样本的概率分布模型很难准确定义，所以，非参数的分类方法应用比较广泛。下面对几种比较成熟的分类方法进行简单的分析和比较：

1．基于外延的分类方法（参数算法）

这种分类方法关心的不是文本的语义内容，而是根据文本的外在特征进行分类。最典型的也是最常见的方法就是基于向量空间模型（Vector Space Module）的方法。这种方法是把文本表征成由特征项构成的向量空间中的一个点，通过计算向量之间的距离，判定文本之间的相似程度。采用该模型的文本分类方法的一般步骤是：先通过对训练语料的学习，对每个类别建立特征向量作为该类别的表征，然后依次计算该向量和各个类别特征向量的距离，选取距离大小符合阈值的类别作为该文本所属的最终类别。这种方法在很多领域得到了广泛的应用，但是其不足之处也是不容忽视的： 一方面，正确率比较低，一般最大能达到 80%左右，而且很难进一步提高[31]；另一方面，对于不同题材的文本，其归类的正确率更是大打折扣。例如，相同的向量空间值，在不同领域的文本分类过程中产生，就会在很大程度上妨碍机器的正确判断能力。

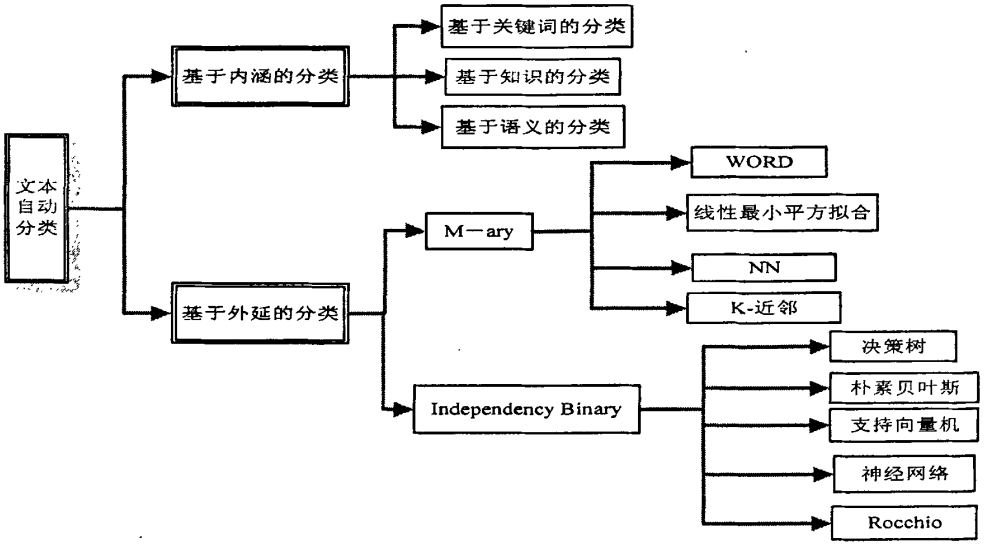

图 5—2 文本自动分类方法

2．基于内涵的分类方法（非参数算法）

顾名思义，这种类型的方法就是从分类文本出发，或者从内容着手，或者从字面着手，对待分类文本进行分析，依据一定的规则或者函数来确定待分类文本所属的类别。主要有以下三种技术：

（1）基于词语的文本分类技术

词是概念的基本构成单位，而语义信息是基于概念之上的。从文本中抽取能够反映该文本内容的关键词，通过对关键词的归类而进行文本分类。很明显利用这种方法产生的分类结果并不能真实的代表全文的语义内容，所以这种方法产生的归类其实也并不是真正通过语义来进行归类，属于文本分类研究早期采用的技术[32]。

（2）基于知识的文本分类技术

基于知识的分类技术必须要有一个明确的知识库，知识的表示方法主要有规则库、语义模型或格框架等。基于知识的文本分类技术最大的特点，也是其不足之处，就是需要用户了解整篇文章的意思，还必须要有领域知识专家来建造一个知识库。这样一来，在某个领域建立知识库就会产生专业性太强，难以移植等困难。不过根据最近的研究工作可以看出，在一定专业领域内，基于知识库的文本分类系统能够对文本进行准确快速的分类[33]。

（3）基于语义的文本分类技术

基于语义的文本分类技术是一种介于基于词语的分类技术和基于知识的分类技术之间的技术。这种方法只是抽取那些对文本分类研究有用的概念，即抽取领域词语或短语潜在的语义概念进行文本类别的确定。另外，基于语义的文本分类技术并不需要研究者或者计算机去全面理解文章的意思，而只是通过概念和语义之间的关联来对文本进行准确快速的分类，这种方法的运用相对于自然语言理解水平尚处于初级层次的现状而言，无疑是一种既简单又实用的好方法，也是最适合本文的分类方法。

5.3 分类过程

语言学文献自动分类能否顺利实现，主要取决于其中若干个阶段的实现。整个分类过程大致可以分为三个步骤：对待分类文本进行预处理（分词和词性标注）、特征选择、本体解析和文本分类（判定文本类别）。由于篇幅有限，下面只介绍对文献《浅谈节目主持人语音不规范现象及对策》[34] 和《反义语素构词的结构和语义考察》[35]的分类过程。

5.3.1 文本预处理

文本预处理的过程就是分词和词性标注的过程。首先，采用已经比较成熟的海量分词软件（海量分词研究版）[36]，在语意计算模式下，对这两篇语言学文献进行分词和词性标注，这里只列举对其中一篇文献进行文本预处理的过程。如图 5－3 所示。

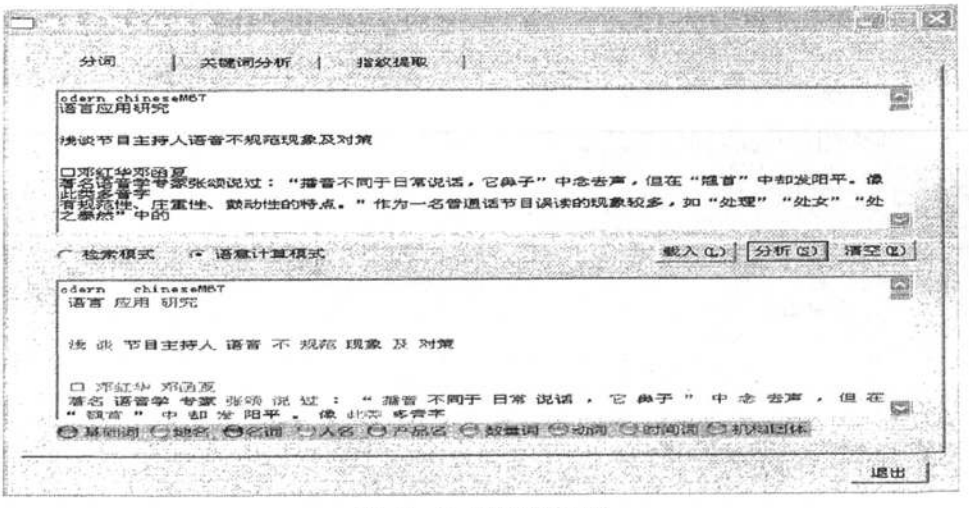

图 5－3 文本预处理

5.3.2 特征选择

可以通过现有的比较成熟的海量分词系统中关键词的分析，利用对权重的计算来确定文本关键词，如图 5－4 所示。

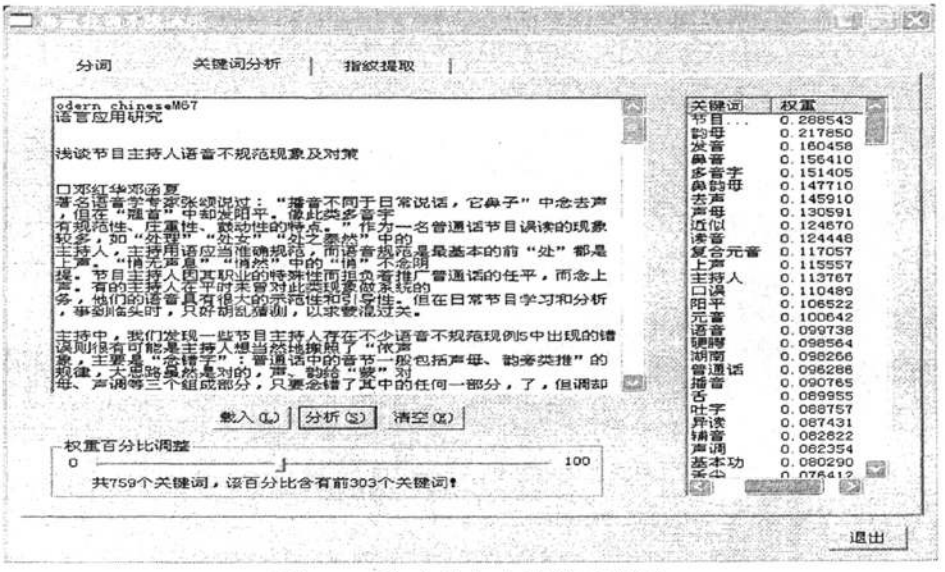

图 5－4 待分类文本关键词分析

5.3.1 中的分词和词性标注的工作，已经为特征选择做了铺垫，在这里，利用分词工具，共分析出 759 个关键词，而选择其中权重比较高的 303 个关键词进行再分析。运用经过一定改进的 TFIDF 算法（第 3 章构建领域本体库的关键技术中，已经作了比较详细的介绍，这里不再赘述），对语言学文献的关键词进行提取，这个过程必须要有领域专家的参与，领域专家凭借掌握的领域知识，对文本关键词进行进一步处理。包括以下四个步骤：

1．首先明确一些已经明显标注词性的虚词，包括介词、连词、助词、叹词等，通过统计文本中每个词出现的频率，预先定义出现频率的最小值，小于这个值的词也可以在特征选择的过程中去掉。

2．在待分类文本中，出现比较多的词语，如“近似”、“主持人”、“湖南”、“播音”等关键词，虽然权重很高，但是对领域知识的表现性却很弱，也要同文本中的虚词等一起去掉。

3．经过上述两个步骤之后，就可以确定那些具有很强类别表示功能的实词作为文本特征项。但是，其中权重较高的“语言学”、“发音”等，在语言学领域只是作为一个

大的聚类存在，可以经过领域专家的确认之后，作为对文献类别表现较弱的词语去掉。

4. 关键词提取过程中，还会出现一些单字词，比如在这篇语言学文献中出现的“舌”字，虽然权重很高，但是就单字词本身来说，经常会存在多个义项，如果作为文本特征项出现，不但会在很大程度上增加运算的负担，而且，由于单字词本身的多义性，作为文本特征项在分类过程中，很有可能会影响文本类别的判定。因此，对于单字词，也需要在领域专家的参与下，对其进行舍弃处理。

经过上述处理，我们就可以确定文献《浅谈节目主持人语音不规范现象及对策》的文本特征项为：“声母、韵母、平声、上声、发音、语音、声调、阴平、口语、调类、声韵、元音”等。用这些经过处理的文本特征项与领域本体库中的领域特征项相匹配，有利于提高匹配工作的准确性。

而同样的，对文献《反义语素构词的结构和语义考察》也进行上述工作，可以确定文献的特征项为：“语素、构词、新词、虚化、引申、委婉、贬义词、褒义词”等。虽然在对文献进行分析的过程中，也产生了一些语法范畴的核心概念，但是经过认真推敲，这些词语毕竟属于少数，而且对文章类别表示意义不是很明显，因此，这里也可以忽略不计。

另外，在本文的语言学领域本体，也是在特征选择的基础上建立起来的，前期工作的一致性，也能够在很大程度上确保后面文本分类过程中匹配的科学性。

5.3.3 本体解析

就语言学本体而言，可以从它所描述的概念与概念之间的关系得到概念之间相对的映射规则[37]，如图 5—5 所示：

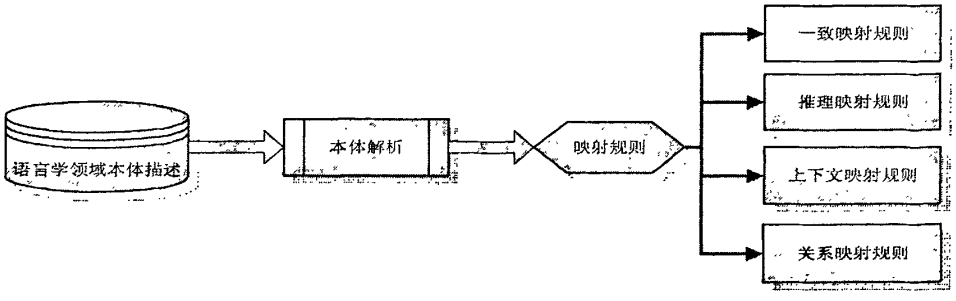

图 5—5 本体解析生成映射规则

1. 一致性映射，是指文本特征项与本体库中的概念、实例相匹配的规则。例如，

在语音学中，“声调”概念对应的是“声调－音调”这一领域特征项。

2．推理映射，是指与概念的正规表达式以及关键词相匹配的规则。例如，“元音”对应的是“音素”这一聚类，而“音素”对应的又是“语音学”的聚类。由此可以把“元音”这一实例映射到“语音学”的聚类中。

3．上下文映射，是指概念之间上下文关系的匹配规则。例如，在语音学本体中，“音素”和“音节”之间存在着上下文关系。

4．关系映射，是与概念之间的结构关系相匹配的规则。例如，作为概念的“声调”与“阴平、阳平、上声、去声”等实例之间存在着结构关系。

利用语言学领域本体产生的四种映射规则对语言学文献进行分类。

1．依据一致性映射规则进行的匹配是精确匹配，也称直接匹配。在语言学领域本体的某一平面确实存在与这一文本特征项相同的领域特征项，依此就可以进行精确匹配。这样的精确匹配在同一篇文献中达到一定的数量，我们就可以由此判定文本所属类别。

2．依据推理映射规则、上下文映射规则和关系映射规则进行的匹配是模糊匹配，也可以称作间接匹配。文本特征项通过本体的推理映射规则、上下文映射规则和关系映射规则，与领域本体中的特征项产生千丝万缕的联系，这样，通过一定的算法，就可以把文本归入某一类别。

5.3.4 基于本体的分类算法

很显然，基于本体的分类方法就是要建立在词语之间语义关联的基础之上。综合考虑 5.2 对几种文本分类方法的分析和比较，基于外延的方法最先排除，而应该选择基于内涵的文本分类方法。前面已经介绍过，基于关键词和基于知识的分类方法，本身都存在很大的缺陷，不能对文本进行准确、科学的分类。因此这里选择基于语义的文本分类方法。

不管选择何种分类方法，一般都不会脱离一个核心思想，也是文本分类的中心任务：匹配。一般的分类都是用待分类文本经过预处理和特征选择以后所形成的模块，和一定的规则或者特征相匹配，计算匹配的相似程度，从而得出结论：待分类模块是否与这一类别模块匹配成功。

文本分类本质上就是一个映射的过程，本体解析中提到的四种基于本体的映射规则，也就是基于本体的文本分类规则。具体到分类过程中，就是待分类文本与领域特征库之间的匹配。匹配算法返回的结果有以下几种可能[38]（C_1表示待分类文本中的特征项，C_2表示领域本体库中的特征项）：

1．Exact。指精确匹配，即C_1和C_2是完全相同的概念。反映在 OWL 中，即指向同一节点。

2．Plugin。指C_1是C_2的子概念，即$C_1 \subseteq C_2$。

3．Subsume。与 Plugin 相反，Subsume 指C_2是C_1的子概念。即$C_1 \supseteq C_2$。

4．Intersection。C_1和C_2之间存在交集。

5．Fail。除上面四种匹配结果之外的结果，都为 Fail，这个结果代表匹配失败。

基于语义的匹配方法有很多种，这里就其中的两种算法来具体分析本体映射规则在分类过程中的应用。这两种算法都以上述 5 种结果为标准来评估匹配是否成功。

1．概念语义的匹配

因为基于概念语义的匹配类型中的概念和匹配对象一般来自领域本体中的概念，所以，在进行概念语义匹配时，可以直接应用领域本体中的概念进行匹配。需要借助领域本体库层次结构中对概念语义关系的描述来计算语义相似度，根据这个数据来判定文本所属的类别。

同样的词语在不同的上下文中会有不同的词义，本文的工作是假设已经进行了词义排歧处理，也就是说，这里的分析都是建立在词语被标注成概念的基础之上，对概念进行语义比较的。本文把相似度的取值范围定义在 0 到 1 之间，当相匹配的两个概念相同时，它们之间的语义相似度为 1；当两个概念语义没有丝毫关联，它们之间的语义相似度为 0，其他情况下，相似度的取值在 0 到 1 之间。而且，相似度越大，证明两个词语越相近；相反地，相似度越小，证明两个词语越相背离。

这里还要再次提到一个概念：语义距离。第 3 章在介绍特征选择算法时已经说过，语义距离是指同一个本体中不同类间关系链中最短关系链的长度。一般来讲，两个概念之间的语义距离越近，语义相似程度越高，反之则越低。当对待分类文本与领域本体进

行匹配时，语义距离数值越小，说明领域本体与待分类文本越接近，当语义距离为 0 时，领域本体可以完全与待分类文本相匹配；当语义距离超过一定数值时，就可以认为领域本体与待分类文本无关，不能作为判定文本属于某个类别的标准。

由此可见，语义距离和相似度的关系是：

（1） 当语义距离为 0 时，相似度为 1；

（2） 相似度随语义距离的增加而减小；

（3） 相似度必须保证在[0，1]区间之内。

根据领域本体的概念层次结构，很容易想到可以用两个概念在结构图中的最短路径距离来表示它们之间的语义关系[39]。设定：两个概念 C_1，C_2，两者之间的语义距离 $Dist(C_1,C_2)$ 为连接它们最短路径上的 n 条边的权值的总和。即：

$$Dist(C_1,C_2) = \sum_{i=1}^{n} weight_i \qquad (1)$$

其中，$weight_i$ 是连接 C_1，C_2 的最短路径上第 i 条边的权值。

一般情况下，不考虑其他任何因素的影响，每条边对语义距离的贡献都是相同的，树中两个节点的最短路径距离就是连接它们的最短路径上边的条数。可以设定每条边上的权值都是 1，即 $weight_i$=1。以对文献《浅谈节目主持人语音不规范现象及对策》进行文本预处理以后得到的特征项为例进行表示，则图 5-6 中：

$Dist$(音节，音素)=2；

$Dist$(声母，韵母，声调) =2；

$Dist$(单元音韵母，复元音韵母，鼻音韵母)=2；

$Dist$(阴平，阳平，上声，去声) =2；

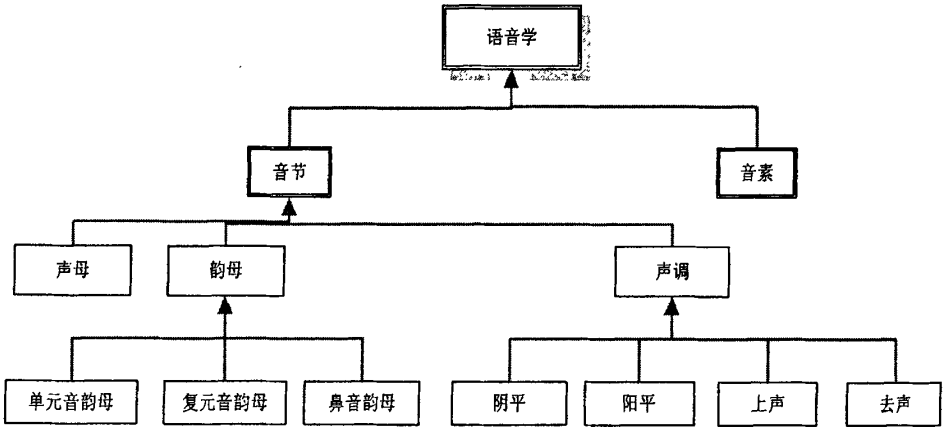

图 5－6 语音学本体中的音节概念层次

由图 5－6 观察可知，影响概念之间语义相似程度大小的音素有以下几个方面：

（1）概念在树中所处的深度

这四组概念的语义距离相等，表明它们的语义相似度是一样的。但是，按照我们的主观判断，显然，后两组概念的语义相似程度明显高于前两对。进一步观察可以发现，处于层次树中距离树根较远的概念之间的相似度要比距离树根较近的概念之间的相似度大一些。这是因为，在概念层次树中，从上到下，概念的分类也是由大到小的，大类别之间的相似度肯定要小于小的类别。所以，概念在树中所处的深度也是必须要考虑的一个影响因素。处于不同深度的边，应该赋予不同的权值。

这里采用基于 SUMO 的概念语义相似度的计算方法[40]，计算概念之间的语义相似度：

首先，概念 C 在层次树结构中的深度 $Depth(C)$ 等于该概念与树根 R 的最短路径中所包含的边数，即

$$Depth(C) = \sum_{i=1}^{n} 1 \qquad (2)$$

其中，n 是该最短路径包括的边数，而 $Depth(R)$ =0；对于一棵树 T 的深度 $Depth(T)=\max(Depth(C))$；其中 C 为树 T 上的任一概念；

第二步，按照本文的定义，所有从概念 C 引出的边具有相等的权值，简称为概念 C 的权值，记做 $weight(C)$；

第三步，用 $parent(C)$ 表示概念 C 的父节点，则

$$weight(C) = \frac{1}{2^{\mathrm{Depth(C)}}} \tag{3}$$

由公式（3）可以看出，概念在树中所处的深度越深，那么以它为父节点，引出的所有边的权值越小。这样就可以保证，具有较大深度的概念之间的语义距离相对较小，同时相似程度也越大。

（2）概念的分类细致程度

从图 5—6 观察可以看出，概念“单元音韵母”和“鼻音韵母”的父节点“韵母”与概念“阴平”和“上声”的父节点“声调”在树中处于同样的深度，根据公式（3）计算，从“韵母”和“声调”引出的边的权值是相等的，再根据公式（2），“单元音韵母”和“鼻音韵母”的语义距离与“阴平”和“上声”的语义距离也是相等的，但是按照人的主观判断，后面两个概念的语义距离较为接近，也就是它们之间的语义距离要小于前面的一对概念。观察发现，它们的区别主要在于后面一对概念的兄弟节点数目相对较大，也就是它们的父节点下分类的细致程度比较高。所以，一个概念的分类细致程度也是影响语义距离，进而影响概念之间语义相似程度的一个重要因素。

用 $Wid(C)$ 表示概念 C 的宽度，即其子节点的数目，可以修改概念 C 的权值为：

$$weight(C) = \frac{1}{Wid(C)} \times \frac{1}{2^{\mathrm{Depth(C)}}} \tag{4}$$

这样，处于相同深度的概念宽度越大，其权值就越低，反之则越高。但是，对于任一概念 C，当 $Wid(parent(C)) \geq 2 \times Wid(C)$ 时，有：

$$Wid(parent(C)) = \frac{1}{Wid(parent(C))} \times \frac{1}{2^{Depth(parent(C))}} \leq \frac{1}{2 \times Wid(C)} \times \frac{1}{2^{Depth(parent(C))}}$$

$$= \frac{1}{Wid(C)} \times \frac{1}{2^{Depth(parent(C))+1}}$$

$$= \frac{1}{Wid(C)} \times \frac{1}{2^{Depth(C)}}$$

$$= weight(C)$$

这样的话就背离了随着深度的增加权值降低的思想，所以，在这里可以把公式修改为：

$$weight(C)=\begin{cases}\dfrac{1}{Wid(C)} & C\text{为父节点}\\ \\ \dfrac{1}{Wid(C)}\times\dfrac{1}{2}\times weight(parent(C)) & C\text{为其他节点}\end{cases} \tag{5}$$

这样的公式就可以保证，随着概念在树中所处的深度的深入，概念的权值逐渐变小。同时，还可以保证随着概念分类从粗糙到细致，概念的权值也是逐渐变小，概念之间的语义距离也随之减小。还应该考虑到，既然任一概念 C 下包含的所有节点都是有其细分得到的，那么对于 C，与其下所包含子孙节点的相似程度都应该比与其兄弟节点之间的相似程度更大。

$$weight(C)+weight(C_1)+weight(C_2)+\cdots+weight(C_n)$$

$$\prec weight(C)+\frac{1}{2}weight(C)+\frac{1}{2^2}weight(C)+\cdots+\frac{1}{2^n}weight(C)$$

$$=weight(C)\prec 4weight(C)\le 2weight(parent(C))$$

也就是 C 和其任意子孙节点的距离都小于和其兄弟节点的距离。另外，根据本文的定义以及公式（2）可以得到，$Dist(C_1,C_2)=Dist(C_2,C_1)$，说明概念之间的语义距离具有对称性。

根据上面对语义距离的定义可以知道，语义距离 $Dist(C_1,C_2)\in$（0，2），而相似度 $sim(C_1,C_2)\in(0,1)$，而且两者应该是减函数的关系，于是可以得到一个简单的语义距离与相似度之间相互转换的公式：

$$Sim(C_1,C_2)=1-\sqrt[t]{\tfrac{1}{2}Dist(C_1,C_2)} \tag{6}$$

其中 t 是一个可调节的参数。

参看图 5—6，对于概念“音节”和“上声”，根据我们的主观判断，“音节”和“上声”相比的相似程度要小于“上声”和“音节”的相似程度，也就是概念间的语义相似

程度具有不完全对称性。对于一个概念而言，子与父的相似程度高于父与子的相似程度。

$$\text{即：}\ Sim(C_1,C_2)\begin{cases} \succ sim(C_2,C_1), & Depth(C_1)\succ Depth(C_2) \\ \leq sim(C_2,C_1), & Depth(C_1)\leq Depth(C_2) \end{cases}$$

综上所述，就可以得到概念之间的语义相似度计算公式：

$$Sim(C_1,C_2)=1-\sqrt[4]{\tfrac{1}{2}\times a\times Dist(C_1,C_2)} \tag{7}$$

其中

$$a=\frac{Depth(C_2)}{Depth(C_1)+Depth(C_2)}$$

对特征提取过程中产生的权重最靠前的 20 个特征项进行上式的计算，可以得到：文献与领域本体库中特征项匹配的相似度明显大于 0.5，而随着参与计算特征项的不断增加，计算出来的相似度更加接近于 1。因此，可以判定文献与语音学领域本体库匹配成功，文献可以划归语音学类别。同理，也可以用这种方法对其他文献进行分类，有了领域本体的参与，分类的准确性有了很大的提高。

2．非一致性模糊匹配

下面介绍另外一种更容易理解的匹配方法：非一致性模糊匹配方法。这里在原有的非一致性模糊匹配方法的基础上，对其进行一定改进，使之更适合应用领域本体库中对语义关系的描述。

从概念就可以知道，非一致性模糊匹配算法弥补了一致性匹配（搜索项必须在数据库中找到原型才能匹配成功）的不足，即使在数据库中没有与用户查询关键词严格匹配的记录，通过本体内部的语义扩展，或者按照领域本体库中记录的与用户查询关键词相近度的计算，同样可以搜索到用户需要的信息，更准确的对所需文本进行类别匹配[41]。

为了描述本算法，首先要明确两个概念：相似度和背离度。

设 $T=t_1,t_2,\ldots,t_n$ 为领域本体内的特征项；以它作为依据去匹配待分类文本的特征项；

设 $S=s_1,s_2,\ldots,s_n$ 是待分类文本经过文本预处理模块和特征选择模块之后产生的特征项。

模糊匹配要求得到的匹配结果是：待分类文本中的特征项具有领域本体库中特征项的基本特征。

基于一般性考虑，对文本特征项的特征做如下分析：

（1）一个特征项的特征由这个关键词本身完全体现（也就是上文本体解析中所说的一致性映射规则），那么，这个特征项与领域本体中的特征项完全相同，是模糊匹配中的特例——精确匹配。

（2）设定领域本体库中特征项集合 T 为 $T_{i,j}=T_iT_{i+1}\ldots T_j$，其中，$1\le i\le j\le n$（很明显 $T=T_{1,n}$）。

如果语言学文本中存在 $S_{1,n}=S_1S_2\ldots S_n$ 在 $T_{1,n}=T_1T_2\ldots T_n$ 中对应位置上包含一系列子串，$T_{k_1,k_2},T_{k_3,k_4},\ldots,T_{k_x,k_y}\ (1\le k_1\le k_2\prec k_3\le k_4\ldots\prec k_x\le k_y\le n)$

待分类文本中的特征集合 S 与领域本体库中的特征集合 T 的相似程度标记为 $sim(S,T)$，

$$sim(S,T)=\sum Len(T_{ki,kg})/Len(T) \qquad (8)$$

其中 Len 表示特征项的长度。

依据本节开头部分对五种匹配结果的分析，可以得出结论：

（1）显然 $sim(S,T)$ =1 时，S =T，即 S 与 T 精确匹配。

设定 s_1 =“声调-音调”，在进行匹配时，发现领域本体库中存在这样的特征项 t_1，与 s_1 完全一致，那么就可以肯定待分类文本特征项 s_1 与领域本体库中的特征项 t_1 精确匹配。与此同时，s_2 =“单元音韵母”，与领域本体中的特征项 t_2 精确匹配……依此类推，如果一篇待分类文本中，类似这样的精确匹配出现频率达到设定的阈值，那么，就判定这篇文章可以归属到语言学领域中的普通语言学，再细致一些就可以归入普通语言学中的语音学范畴。

（2）S 与 T 的背离度记为 $dif(S,T)=1-sim(S,T)$，当 $sim(S,T)$ 接近 1，$dif(S,T)$ 接近 0 时，可以认为 S 与 T 模糊匹配，也就是间接匹配，可以大致确定文本的类别。当 $sim(S,T)$ 接近 0，而相对的 $dif(S,T)$ 接近 1 时，文本则不能与本体库中的特征项匹配，

不能归入此类别。

设定s_3 ="鼻音韵母"; t_3 ="单元音韵母", s_3与t_3是一个概念领属下的两个互相独立的属性，通过计算可以看出，两个词语之间是模糊匹配的关系，而且通过计算它们之间的语义相似度，$sim(s_3,t_3)$接近 1。以此类推，这样的模糊匹配出现的次数达到一定的阈值，就可以判定待分类文本属于该类别。

如果待分类文本特征项s_4 ="双宾语句"，通过以上公式计算与领域本体库中特征项之间$sim(s_4,t_4)$接近 0，而$dif(s_4,t_4)$接近 1，就可以判定这个文本不能与领域本体库中的特征项相匹配，因此不能归入其中。

相似度或背离度为模糊匹配提供了判定成功与否的依据，并可以通过设定相似度或背离度的值来控制模糊匹配的结果[42]。根据以上的分析，可以得出以下结论:

（1）相似度的计算遵循顺序性，即领域本体库中包含的特征项保持位置上的对应。

（2）领域本体库中特征项所包含的义项越多，则用待分类文本中的特征项进行匹配时相似程度就越大。

（3）单个特征项只能代表领域本体的一部分特征，尤其特征项表示很长时，单个特征项不会对整个领域本体的特征产生很大的影响。这也是模糊匹配的基础。

（4）对于不匹配的特征项，其长度越大，背离度就越大。当背离度超过某阈值时，可以放弃当前的匹配。反之，若匹配后背离度没有超过设定的阈值，则可认为模糊匹配成功。

这是两种比较简单的，便于实施的文本分类方法。在基于本体的语言学文本分类过程中具有较强的可操作性。通过这样的计算过程，可以得到分类结果:《浅谈节目主持人语音不规范现象及对策》属于语音学领域文献,《反义语素构词的结构和语义考察》属于词汇学领域文献。这样就可以把以往常用的文本分类工具不能精确分类的文献，比较准确的划分类别，分别归入其所属的学科：语音学、词汇学、语法学、方言学、修辞学等。在分类过程中，有了领域本体的支持更凸现分类的准确性和科学性，从而很好的克服了以往分类工具的粗线条分类模式。使文本分类更加细致、精确。

5.4 分类结果评估

对待分类的 50 篇普通语言学文献运用比较常用的文本分类工具：TRS 文本挖掘基础件(TRS CKM)[43]在复杂模板下进行分类，下面以《浅谈节目主持人语音不规范现象及对策》和《反义语素构词的结构和语义考察》为例，得到分类结果列表表示如下：

表 5－1　TRS 分类结果演示

《浅谈节目主持人语音不规范现象及对策》

分类结果：	文化事业 005\文学艺术 005002
摘要：	有的主持人在主持节目的过程中过于随意，或希望营造一种轻松的氛围，说话时不太注意自己的发音。由于情绪亢奋，说话速度过快；或是心理过度紧张，发音器官未打开，从而造成发音不到位。
关键字：	发音 音变 口语等

由上表可以看出，TRS 文本分类系统对文本的类别归属判定为文化事业中的文学艺术类。那是由于 TRS 文本分类系统没有对领域内的诸多语义关系进行很准确的描述和界定，导致分类结果只能停留在大的类别上，而未能对文本进行更细致的分类。同样，对另外一篇文献《反义语素构词的结构和语义考察》的分类也是如此。如下表所示：

表 5－2　TRS 分类结果演示

《反义语素构词的结构和语义考察》

分类结果：	文化事业 005\文学艺术 005002
摘要：	反义语素构词的结构和语义考察。□吴建勇[摘要]反义语素构词是一个特殊的语言现象。本文就反义语素构词在结构、语序和意义等方面考察其具有的。特点，以便更好地掌握和运用这类词。日常生活中我们经常会遇到这样一些词，如：“黑”和“白”是一对反义语素，可它们却能组合在一起构成新词“黑白”而且具有新的词义。汉语中这种由两个具有相反意义的语素构成的词语还有很多，如：迟早、先后、昼夜。这些词很有特点，尽管这些构词语素在意义上对立，但它们却能组合在一起。考察反义语素构词的结构特点，对于我们更好地掌握和运用这些词大有帮助。
关键字：	语素 成分 构成

在对这篇文章的分类过程中，TRS 分类工具对关键词和摘要的提取都比较准确，但是，把文章归为文化事业中的文学艺术类别，这样的分类过于粗疏。同时也暴露了分类工具明显的缺点和不足。

对 50 篇语言学示例文献，运用该文本分类工具进行分类结果与上述两篇文章的分类结果大致相同，都没有能够对文本类别进行细致、准确的归属。

本文在加入本体论思想的文本分类方法中，先对待分类文本，经过文本预处理和特征选择过程后，产生文本特征项与领域本体库中的特征项相匹配，计算文本相似度，并通过此相似度的大小来判定文本所属类别。利用上述过程进行计算，得到匹配是否成功的结果，从而获得分类结果。同时，以人工分类结果作为参考，判断分类方法的优劣。

在我们选择的 50 篇普通语言学文献中，运用上述两种基于本体的分类方法分别对该文本与本体的相似度进行计算，可以划归到“普通语言学”类别，或者更准确一些，可以划归到“语音学”、“词汇学”、“语法学”、“方言学”等类别的结果准确率均达到 80%以上。由此可见，基于本体的文本分类方法明显优于传统的分类方法。

同时，还会有一些分类不准确的问题不容忽视。如对《山东方言区普通话水平测试各等级语音面貌分析》进行分类时，结果产生了一些误差，对文章进行特征提取，结果产生了一些权重较高的特征项“重音”、“语调”、“轻音”等，把这些关键词与领域本体库相匹配，得到文章类别属于语音学领域。但是，实际上经过领域专家分析，这篇文章应该属于方言学范畴。正是因为文章主题往往表现出不同的类别特征，很容易在操作过程中产生误差，影响类别的判定。这就需要我们对领域特征项的选择和领域本体的建立做更深入的研究，使之不断完善，努力提高分类的准确率。

分类结果的评估需要以人工分类的结果（假设正确）为参照。评估得到的结果还要对领域本体库进行反馈，利用匹配成功的特征项和匹配失败的特征项进一步完善本体，补充概念实例和概念关系。

5.5 本章小结

本章提出了一种基于本体的文本分类方法，介绍了两种匹配算法：基于概念语义的匹配和非一致性模糊匹配。分别用这两种分类方法对语言学文献进行自动分类，传统的文本分类方法不能较好的对语言学文献进行准确的分类，在对领域特征项加入语义关

系之后建构出来的领域本体库的基础上，可以部分地克服传统方法的缺陷，更多地把语义方面的关联考虑进来，提高文本分类的准确率。

第 6 章　结论和展望

6.1 结论

随着网络的日益普及，文本信息迅速膨胀，使得文本分类技术成为信息技术领域的一个重要研究内容。语言学作为一个集多门边缘学科为一体的古老而又年轻的领域，对其文献进行准确的检索和分类成为摆在研究者面前的一项重要课题。传统的分类方法只是限于表层的分类，无法对语言学文献本身所包含的各项语义信息进行很好地理解和运用。把本体论的思想融入语言学文献的文本分类过程无疑是一种有益的尝试。本文的工作如下：

1．介绍文本分类技术在国内外的研究现状，分析了语言学文献的特点，指出了对语言学文献进行分类研究的重要意义。

2．从理论上对本体的知识作了详细的介绍，阐明了本体的来源、定义、类别、构建方法等。

3．介绍了构建领域本体所需要的几项关键技术，首先采用向量空间模型的方法表示文本；采用经过一定改进的 TFIDF 算法确定文本特征项，并在领域专家的参与下，对其进行语义扩展，形成领域本体的特征项。

4．提出了在语言学领域建构领域本体的全过程。首先确定领域本体的范畴和目的；在此基础上，运用 Protégé系统，建立了语言学领域的本体库。

5．介绍了基于本体的语言学文本分类流程，简要分析了几种文本分类方法，并最终选择基于概念语义的分类方法和非一致性模糊匹配算法对文本进行分类。采用海量分词系统对文本进行分词和词性标注，然后抽取关键词。并在此基础上确定文本特征项，与领域本体库相匹配，运用上述两种方法，通过相似度的计算最终确定文本所属类别。

6.2 展望

本文提出一种基于本体的语言学文献自动分类方法，在本体的辅助下，将背景知识整合到分类领域中，随着背景知识的引入，实例向量表达了更为丰富的涵义，使得学习

能够更加准确。在示例中，本体是与分类原则紧密相关的。在使用过程中，应该将本体限定在与分类规则相关的领域本体中。分类的好坏依赖于本体描述分类规则相关领域的能力。随着数据库的规范，主题词表的不断完善，领域本体库也将更趋完整，系统自动分词和分类的功能也会更好，效率也会越来越高。与此同时，还会开拓一些新的研究领域，例如：

1．本体对于不同领域的运用，以及如何运用；

2．应该建立怎样的相关本体，领域本体之间的关系如何确定；

3．本体如何共享，怎样在一个领域建立本体以后可以在很多不同领域发挥作用。

参考文献

[1] 曾伏虎， 曹焕光， 曹素青. 一个中文文本自动分类数学模型. 情报学报 1999. (1). 18-19.

[2] F. Sebastiani. Machine Learning in Automated Text Categorization. ACM Computing Surveys. 2002. 34(1). 1-47.

[3] T. M. Mitchell. Tom M. Machine Learning. McGraw-Hill Education, Inc. 2003. 42-47.

[4] T. M. Cover, P. E. Hart. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory. 1967. 13(1). 21-27.

[5] 靳小波. 文本分类综述. 自动化博览，2006. 11（23）. 24-29.

[6] N. J. Belkin, W. B. Croft. Information Filtering and Information Retrieval: Two Sides of a Same Com-munication of ACM, 1992. (35)12. 29-37.

[7] 肖琨焘，李德顺. 本体. 中国大百科全书，1987（哲学卷 1）. 35.

[8] 耿科明. 基于本体的个性化信息服务研究. 河北大学硕士论文，2006.

[9] T. Ontolingua: A translation approach to portable ontology specifications Knowledge Acquisition, 1993. 5(2). 199-200.

[10] J .Arpirez, A. G. Perez, A .Lozano, et al. (Onto) 2Agent: An Ontology based WWW Broker to Select Ontologies. In: A. Gomez-Perez, VR. Benjamin. Eds. Proceedings of the Workshop on Application of Ontologies and Problem Solving Methods UK, 1998.

[11] 蔡自兴，徐光佑. 人工智能及其应用(第二版). 清华大学出版社. 1996. 121-123.

[12] 常毅，张鑫，基于关键词表达式模型的文本自动分类系统的研究与实现，中国科学院计算机研究所.

[13] 黄萱菁，夏迎炬，吴立德. 基于向量空间模型的文本过滤系统. 软件学报. 2003. (3). 11.

[14] R. O. Duda, P. E. Hart. Pattern Classification and Scene Analysis. New York: John Wiley and Sons, 1973.

[15] 王梦云，王素格. 一个基于字特征的文本分类模型. 计算机工程与应用. 2004. (13). 64-65.

[16] 王洋，秦兵，郑实福. 句子相似度计算在 FAQ 中的应用. 第一届计算语言学学生研讨会论文集. 北京:北京大学. 第一届学生计算语言学研讨会. 2002. 175-181.

[17] L. A. Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including owl. In Proceedings of the international conference on Knowledge capture, ACM Press. 2003. 121-128.

[18] T. B. Lee, J. Hendler, O. Lassila. The semantic web. Scientific American, 2001. 284 (5). 34-43.

[19] The Protégé project, http://Protégé.stanford.edu, 2005.

[20] H. Matthew, K. Holger, A. Rector. A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0[EB/OL].http://www.co-ode.org, 2004. 8.

[21] 董慧，余传明，杨宁. 基于本体的数字图书馆检索模型研究（III）历史领域资源本体构建. 武汉大学信息资源研究中心. 情报学报. 2006. 4. 45-46.

[22] 林春梅，何跃. 创建企业本体模型的研究，微机发展，2003.（1）. 23- 25.

[23] 李景，苏晓鹭，钱平. 构建领域本体的方法. 计算机与农业. 2003. （7）. 45-47.

[24] 杜小勇，李曼，王大治. 语义 Web 与本体研究综述. 计算机应用，2004. （10）. 29-32.

[25] 林春梅，金鑫，何跃. 基于 RDF 构建主义化本体模型，计算机应用与软件. 2002. (6). 32-40.

[26] 许珏. 本体论与信息检索. 中国信息导报 2004. (3). 57-58.

[27] 赵国涛，何钦铭. 基于本体的异构文本分类系统 计算机工程 2004. 11. 第 30 卷(21). 123.

[28] D. David. Lewis, et al. Training Algorithms for Linear Text Classifiers.SI GI R96:Proceedings of the 19th Annusl Intmational ACM-SIGIR Comference, Konstanz: Hartung- Gorre Verlag, 1996. 298-306.

[29] 吴赣，程学旗，余智华. WWW 页面的文档分类技术. 计算语言学文集. 1999. 10. 34-36.

[30] W. B. Croft, D. J. Harper. Using probabilistic models of document retrieval without relevance feedback. Journal of Documentation. 1979. 35(4). 285-295.

[31] 苏伟峰，李绍滋. 一个基于概念的中文文本分类模型. 厦门大学硕士论文. 2004.

[32] 曹素丽，曾伏虎，曹焕光. 基于汉字字频的中文文本自动分类系统. 山西大学学报. 1999. 2. 12-15.

[33] 杨建武. 文本挖掘技术 北京大学计算机科学技术研究所. 教学讲义.

[34] 邓红华，邓函夏. 浅谈节目主持人语音不规范现象及对策. 现代语文(语言研究版). 2006. 12. 67-68.

[35] 吴建勇. 反义语素构词的结构和语义考察. 现代语文（语言研究版）. 2006. 4. 37-38.

[36] 海量分词软件介绍. http://www.hylanda.com.

[37] 马征. 基于本体的 Web 页面分类挖掘. 中南大学硕士论文. 2004. 5

[38] 罗洋，曾国荪. 基于本体语义的网格服务能力匹配算法. 计算机应用. 2004. 24(9). 52-54.

[39] I. Niles, A. Pease. Towards a Standard Upper Ontology. Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).Maine, 2001. 2 - 9.

[40] 徐德蕾，郑春卉，K. Passi. 基于 SUMO 的概念语义相似度研究. 计算机应用. 2006. (1). 26.

[41] 李名标. 模糊分类与模糊匹配结合的模糊检索. 计算机科学. 2000. 11(8). 37-39.

[42] 潘景昌，孙玉辉，徐义明. 一种简易的模糊匹配算法的实现. 信号技术与信息化. 2006. (3). 131-132.

[43] TRS 文本挖掘基础件(TRS CKM)介绍. http://www.trs.com.cn/products/textmine/trsckm/.

攻读硕士学位期间发表论文情况

曹亚妹. “有”字句的主语和宾语的自动界定. 江汉论坛，2006.4.153-155.

致 谢

值此论文完成之际，向我的导师田学东教授表示深深的谢意。本文从选题、撰写到完成都凝聚了导师大量的心血和汗水。在此衷心感谢田老师的悉心指导和教诲！没有导师的帮助，就没有我今天的成果，他严谨治学的作风和平易近人的态度将是我终生学习的榜样。

另外，感谢王强军老师和李新福老师对我学习和研究上的指导与帮助。

田学东教授和王强军老师早在论文选题之初，就针对涉及的知识和技术问题给予了多方面的指导，使得此论文的资料准备工作得以顺利进行，避免了错误方向造成的时间和精力的浪费。

在研究工作的逐步展开工程中，田学东老师、王强军老师和李新福老师经常组织我们讨论研究的方向和实现的细节，在三位老师的严格要求和悉心指导下，研究工作顺利进行直至论文的如期完成。

在本论文的完成过程中，得到了同一研究课题中张自儒、赵晓华、宋丽娟、杜娟等同学的大力帮助，在此对他们表示衷心的感谢！

还要感谢我的父母，有了他们的支持我才完成了今天的学业。

感谢对论文进行评审并提出宝贵意见的各位老师和专家！