附赠
完整版电子光盘

汉语科技词系统

（新能源汽车卷）

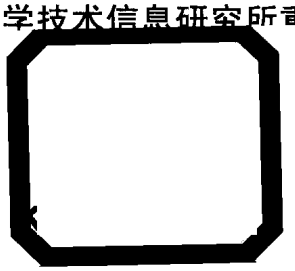
CHINESE SCIENTIFIC & TECHNICAL VOCABULARY SYSTEM

NEW ENERGY VEHICLES

中国科学技术信息研究所　编著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

“十一五”国家科技支撑计划项目“知识组织系统的集成及服务体系研究与实现”（2006BAH03B03）
中国科学技术信息研究所重点工作项目“汉语科技词系统建设与应用工程”（2008 -3-2，ZD-2010-3-2，ZD2011-3-2）



汉语科技词系统

（新能源汽车卷）

中国科学技术信息研究所 编著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

图书在版编目(CIP)数据

汉语科技词系统. 新能源汽车卷 / 中国科学技术信息研究所编著. -北京：科学技术文献出版社，2012.1
ISBN 978-7-5023-7116-6

Ⅰ. ①汉… Ⅱ. ①中… Ⅲ. ①汉语-科技情报-机器检索：情报检索-检索系统 ②新能源-汽车-科技情报-机器检索：情报检索-检索系统 Ⅳ. ①G354.43

中国版本图书馆 CIP 数据核字(2012)第 250863 号

内 容 简 介

本书系统介绍了汉语科技词系统的产生背景、历史及发展概况，阐释了汉语科技词系统的数据模型、建设流程以及新能源汽车汉语科技词系统的建设实践及应用探索过程，并对其中的经验教训进行了总结。以实例形式展示了新能源汽车领域汉语科技词系统的部分建设成果，以方便读者直观了解和使用。

本书主要面向希望了解、学习和应用汉语科技词系统的研究者和学习者，既可以作为新能源汽车领域信息分析和信息处理相关人员的工具书和资源使用手册，也可以作为高校信息管理、情报学、图书馆学等专业本科和研究生的教学参考书，还可以作为各类与知识组织、知识工程、新能源汽车领域相关的机构部门的信息工作者和有关管理人员的学习参考资料。

汉语科技词系统（新能源汽车卷）

策划编辑：周国臻　责任编辑：马　帅　周国臻　责任校对：赵文珍　责任出版：王杰馨

出 版 者　科学技术文献出版社
地　　址　北京市复兴路 15 号　邮编　100038
编 务 部　(010)58882938，58882087(传真)
发 行 部　(010)58882868，58882866（传真）
邮 购 部　(010)58882873
网　　址　http://www.stdp.com.cn
淘宝旗舰店　http://stbook.taobao.com
发 行 者　科学技术文献出版社发行　全国各地新华书店经销
印 刷 者　北京时尚印佳彩色印刷有限公司
版　　次　2012 年 1 月第 1 版　2012 年 1 月第 1 次印刷
开　　本　787×1092　1/16 开
字　　数　646 千
印　　张　29.5
书　　号　ISBN 978-7-5023-7116-6
定　　价　158.00 元（附 1 张 CD）

前　言

在科技创新活动过程中，科研人员选题策划、立项评审、研发实施、结题验收、成果传播和成果评价都需要科技信息的支撑。面对大规模增长的科技信息资源，科研人员对及时、准确地获取信息的需求日益强烈，也对科技信息工作提出了新的、更高的要求，信息资源加工、组织和服务的理论方法以及技术亟需创新。

科技信息工作的发展离不开信息技术尤其是知识技术的进步。知识组织系统是知识技术的核心，是提高各类信息资源开发利用效率的重要工具，也是推进知识服务的关键基础。知识组织系统包括图书分类法、主题词表、本体、词系统等多种类型。中国科学技术信息研究所（以下简称“中信所”）作为科技部直属的国家级科技信息机构，在知识组织系统研究和建设的不同历史阶段做了大量工作，曾先后牵头组织编制了《汉语主题词表》和《电子政务主题词表》，并主持制定了《汉语叙词表编制规则》（GB/T 13190—1991）、《文献多语种叙词表编制规则》（GB/T 15417—1994）、《电子政务主题词表编制规则》（GB/T19486—2004）等国家标准。

当前，海量的网络化数字化信息资源使得知识组织系统产生了强烈变革需求，为了适应更加深入的语义应用和大规模信息即时处理的需求，知识组织系统不但更注重实用化和针对性，而且在某些局部变得日益丰富细化。中信所正是在这样的背景下，进一步加强了知识组织系统相关研究工作，针对科研活动跨学科特点，提出以领域为视角构建全面反映各学科知识相互渗透、交叉融合现状的汉语科技词系统。这是一种新型的知识组织系统，其本质是通过概念、概念内涵和概念之间的关系来构建领域知识体系，并通过相应算法对海量科技信息进行语义分析，实现科技信息的深度加工与服务。

汉语科技词系统从萌芽到发展，从理论到实践，前后历经五载。在研究开发过程中，中国科学技术信息研究所及其合作单位做了大量的研究论证和

试验，努力解决了汉语科技词系统的理论研究、工具开发等环节中的一些关键性问题，包括领域概念知识的整理加工、概念知识的构建、管理和服务平台系统搭建等。2010 年起，中国科学技术信息研究所以新能源汽车领域作为切入点，联合机械工业信息研究院等单位尝试大规模构建汉语科技词系统内容。同时，在汉语科技词系统的基础上还开展了科技监测与评价、专利内容深度分析、企业知识管理和移动知识服务等方面的应用实践。在研究实践过程中，中信所也积累了一定的经验教训，希望它们能够为今后领域汉语科技词系统和其他类型的知识组织系统建设应用提供借鉴。

《汉语科技词系统（新能源汽车卷）》有纸质版和电子版。本书即纸质版主要包括三个部分，第一部分是对汉语科技词系统的相关介绍，便于读者初步了解汉语科技词系统。第二部分是领域汉语科技词系统中的部分实例，实例包含核心词条的名称、基本信息（拼音和英文对译）、定义、概念描述、分类信息、词条属性和词条关系等，以方便读者直观深入了解和进一步利用汉语科技词系统。第三部分是附录，包括收录实例词条的音序索引、笔画索引、ISTIC-NEV 分类法、全部的核心词列表、概念描述节点文字解释以及定义来源参考资料等。汉语科技词系统是一个不断进化的知识组织系统，其知识数量和知识内容也在不断地发展变化。截止 2012 年 3 月，新能源汽车领域汉语科技词系统中包含 55958 条词条，其中 6117 条为核心词，其余为基础词；设计并使用有 78 种二级关系类型，并建有 57164 个关系实例；设计并使用有 45 种二级属性类型，并建有 18309 个属性实例。面向新能源汽车的 NEV 分类法有 4 层 180 个类目，并且构建了 5656 个类目实例。每一个核心词都包含对应的英文译文，系统中还包含 5873 条定义，以及 10260 个拥有 HNC 概念描述的词条（包括了所有的核心词和部分重要基础词）。由于内容体量较大，本书只收录了一部分内容，完整的电子版内容可以从配套光盘或者官方网站（www.vocgrid.org）上获取。

为保证对本书中概念理解的一致性，对于必要的核心词我们会附加定义解释。在这一过程中，编写组参考了《中国大百科全书》、《电动汽车术语》、《机械工程名词》、《交通大辞典》、《汽车构造》、《先进电动汽车技术》等领

域相关的工具书、标准、教科书等，以保证定义的准确性和权威性，在此一并向这些文献的作者、编者和出版单位表示诚挚的感谢。

《汉语科技词系统（新能源汽车卷）》的编写工作是一项非常艰巨的工作，在本书出版之际，感谢在汉语科技词系统研发和内容建设中付出大量心血的科研人员，同时也感谢一直关心和支持这项工作的同行专家。由于时间有限，工作量大，虽然几经审校，但编写过程中不妥和错误之处在所难免，欢迎同行专家和广大读者批评指正，以便我们在后续的更新版本中修正、不断完善。

衷心希望本书的出版能够推动其他领域词系统以及知识组织系统相关研究和事件的发展。汉语科技词系统目前还是一棵小小的幼苗，《汉语科技词系统（新能源汽车卷）》是它的第一片嫩叶，希望在国内外同行专家的帮助和指点下，在不远的将来能够成长为一棵参天大树。

中国科学技术信息研究所所长 贺德方

目　录

第一部分　词系统介绍 1

1　汉语科技词系统的产生背景、历史及发展 3

2　汉语科技词系统的数据模型（知识结构） 4

2.1　总述 4
2.2　词条之间的关系 5
2.3　词条的属性 6
2.4　词条的多维分类 7
2.5　词条的定义 8
2.6　形式化概念描述 8

3　词系统建设 11

3.1　知识架构设计 11
3.2　建设主体及建设流程 12
3.3　协同构建平台开发 13
3.4　状态控制 15
3.5　权限管理及任务划分 16
3.6　辅助构建 17
3.7　可逆和继承 18

4　新能源汽车词系统 20

4.1　领域选择依据 20
4.2　规模及版本 20
4.3　关系类型设置 20
4.4　属性类型设置 21
4.5　新能源汽车词系统的说明 22

5　汉语科技词系统的使用 23

第二部分　新能源汽车词系统实例 27

1　格式说明 29

2　实例正文 31

第三部分　附　录 339

A　实例词条音序索引 341
B　实例词条笔画索引 346
C　ISTIC-NEV 分类法 351
D　全部核心词列表（6117 条） 354
E　范畴及词族索引 409
F　概念描述节点文字解释 452
G　示例中定义来源参考文献 456

后　记 459

第一部分

词系统介绍

1　汉语科技词系统的产生背景、历史及发展

中国科学技术信息研究所（以下简称“中信所”）在知识组织系统建设方面有一定的积累。如我国第一部大型综合性主题词表《汉语主题词表》就是在“七四八工程”支持下，由中信所主持编纂的，该词表曾获国家科技进步二等奖，此外，具有代表性的工作还包括《综合电子政务主题词表》和《中国图书资料分类法》等。但是相比国际上对知识组织系统的持续投入和研究的不断深化发展，我国的知识组织系统建设及应用仍相对落后。为此，中信所立足已有基础，面向实际应用，提出“《汉语主题词表》升级改造工程”，并作为重点工作之一，列入所“18866 工程”。随着调研的深入和研究的进展，此项工作的一部分逐步变更为“汉语科技词系统建设与应用工程”。“汉语科技词系统”在继承叙词表主要知识架构的基础上进行扩充，使之既能满足当前需要，又可以较为方便地向本体转化。每个领域词系统的建设都涉及多个学科，建成的词系统则面向信息检索、知识导航、知识管理以及自动化的信息分析处理等多个方面，不但为信息资源管理本行业提供了服务，而且也为其他行业的应用奠定了领域语义资源基础。

研究团队在对国内外知识组织系统调研的基础上，首次提出“汉语科技词系统”的概念，并对其不断深入研究细化、发展完善，目前理论层面已经基本稳定，协同构建平台也有了一定基础，基本能够满足建设的需求，随着示范性新能源汽车汉语科技词系统的建设实践，到 2010 年，新能源汽车领域汉语科技词系统已基本建成。

汉语科技词系统的发展经历了 3 个主要发展阶段。第一阶段（2007—2008 年）是集成融合的探索阶段，这一阶段重点解决多部词表在形式上的集成融合问题，主要提供概念索引、集成展示等功能，初步将一个领域有关的词表集成在一起，形成了一个词表建设和管理的平台；第二阶段（2008—2009 年）是理论研究阶段，在第一阶段基础上，提供更多的语义信息，对关系类型进行了扩展，形成了一套关系扩展的方法，将关系和属性分离，设计了属性类型，并增加了定义知识，形成了词系统的主要知识结构；第三阶段（2009—2010 年）是实践阶段，开发了词系统协同构建平台，在新能源汽车领域展开建设示范。目前汉语科技词系统正在向两个主要发展方向进军，一是基于已有新能源汽车领域词系统基础上的自动构建方法研究；二是开放共享的免费词表构建服务推动，这两个方向分别代表了人工智能和人本计算的思想。

“汉语科技词系统”最初冠名“汉语”有两个原因：一是构建的知识组织系统主要是用汉语表述的，服务于汉语为母语的相关人员的科技信息资源处理分析目的；二是注重继承已有《汉语主题词表》的财富。但是随着建设实践的发展，可以不局限于汉语资源和汉语用户。由于建设的词系统是面向特定应用领域的，因此又称为“领域科技词系统”，同时，由于服务的范围也可以突破科技范围，所以有时又称为“领域词系统”或者“词系统”。因此“汉语科技词系统”、“领域科技词系统”、“领域词系统”、“词系统”可以通用。

2 汉语科技词系统的数据模型（知识结构）

2.1 总述

吸收了叙词表、词典和本体等知识组织系统的数据模型及设计思想而确定的领域词系统的知识结构，主要包括几个方面：1）词条基本信息；2）词条定义及注释知识；3）词条之间的关系知识；4）词条的属性知识；5）词条的多维分类知识；6）词条形式化概念描述知识。

其中，词条的基本信息包含词条的中文词形、对应的英文翻译、对应的拼音、词汇类型（即核心词/基础词区分）等知识要素。词条的定义主要针对核心词，也就是那些在一个领域中处于核心骨干地位的词条，定义通常来自教科书、百科全书、科技期刊文献和互联网等有关文献，定义可以有不止一条。此外，还可以添加有关的各类型编辑、变更和历史注释。词条之间的关系从宏观讲仍然是等同、层级和相关关系，但是对以上关系类型做了细化，尤其是对相关关系。细化既有通用的部分，也有针对新能源汽车特定的部分。词条的属性是领域词系统新增的，用来表征一些依附于主体存在的属性和属性的具体值，从而更全面地描述词汇（或者概念）。词条的分类是另外一种重要的知识，即从宏观管理角度，通过一个范畴或者粗分类表来管理词汇。此外，从分类主题一体化的角度，还需要给出词汇与真实文本分类的相关关系。系统支持多维分类，主要包括《中国图书馆分类法》（CLC）和《国际专利分类法》（IPC）以及团队自己研制的针对新能源汽车的分类法。词条的形式化概念描述是为了便于计算机处理的尝试，目前采用概念层次网络（HNC）的概念符号体系描述。

总体来看，词汇组织有两种主要的方式，一种方式是明确每个词汇的定义，如果每个词汇的内涵、外延及有关信息都界定得非常清晰，则词与词之间的关系和区别也就不言自明，这就是我们日常使用的词典进行词汇组织的方式。另外一种方式是并不给出词的确定定义，而是描述词与词之间的关系，如果关系描述完备了，则其内涵和外延也就清晰了。当然无论采用哪种方式，都不可能做到对词汇的完备描述。但是一定程度的词汇描述，在实际工作中就可以发挥积极作用。从利用的角度看，计算机处理定义较为困难，而处理关系则比较容易；人易于把握定义，但是对于繁多的关系处理起来比较困难。为了同时满足用户对词汇知识和信息检索的双重需求，对于词汇的组织就应该吸收这两种方式的长处，形成计算机易于处理，且用户易于使用的词汇组织方式。汉语科技词系统既采用了关系、分类等叙词表常用的词汇组织方式，又借鉴了本体关于属性描述的优点。同时，汉语科技词系统也吸收了词典的思想，收入词条定义，并在此基础上更进一步以形式化概念的形式对词条本身做描述，这样计算机处理起来将更为便利。

现有的基于词间关系的词汇组织方式主要是对词汇或词汇后面蕴含的概念进行组

织，从而形成一定的结构。常见的结构有 3 种类型：点集结构、树形结构和网状结构。点集结构就是将同义词、近义词放置到一个集合当中，每个点都代表一个词，词之间聚合进而形成多个点集，点集与点集之间是松散的，即任意两个点集之间没有关联。同义词词典就是以这种方式来组织的，这种组织方式获取相对容易，词间关系描述比较简单。这种组织方式对于信息检索的扩展检索有一定效果，但是对非常重要的上下位关系等都舍弃了。《同义词词林》和 WordNet 的结构是发展了的点集结构，在点集基础上附加了不同约束程度的概念树。

树形结构就是层级结构，可以清晰表现词汇的上下位关系，但是通常构造一个词汇量非常大的树形结构是不现实的，因为虽然很多词汇之间的确存在上下位关系，但是在一个词汇量巨大的系统中，词的意义总是有所交织的，所以很难把每个词汇都能放在词汇树的适当位置。一种解决方案是建立概念的树形关系，然后将与概念有关的词捆绑到对应的概念上，形成一个两层的结构，其中概念层是树形结构。通常在概念层还有其他的关系附加，从而形成了一个树形结构为主的网状结构。常见的就是叙词表，概念层用叙词来表示概念，通过用代关系将词汇层映射到概念层。属分关系就是以层次结构来表示，而参照关系则用其他关系来表达，从而形成网状结构。概念地图也是以这种方式来组织词汇的，只不过概念之间的关系相比叙词表更宽泛。

网状结构是对词汇相关或者参照知识的一种组织方式——维基百科（Wiki）的组织就是这种模式，将和本词条知识有关的其他重要词条通过超链接来揭示，所有的词条形成一个网状结构。在一些类似的知识元服务中，也是采取这样的结构，当然具体的关联规则可能和 Wiki 有所不同。网状结构不强调上下位关系，而是认为它们和其他关系的重要程度相类似，着重揭示不同的词汇所对应的所有的相关词汇。此外，还有一种类似的解决方案，并不构造全部的概念，而是从基本单元角度出发，构造一个概念基本单元树形结构，然后利用这些基本单元的组合来表达词汇，组合可能是简单组合，也可能是很复杂的组合，这种结构的词汇组织的两个代表都发源于中国，都是从汉语的特点出发的。其中一个是知网（HowNet），另外一个是概念层次网络（HNC）理论。

定义和词间关系用于组织词汇各有所长，为此汉语科技词系统将两者结合起来，形成包含集成树形结构、网状结构和点集结构的综合组织方式，并明确了需要描述的 6 类知识，从而在充分利用已有资源基础上解决大规模词汇的组织问题。

2.2 词条之间的关系

在从叙词表到本体的众多类型的知识组织系统中，词条之间的关系都作为一种重要的知识而存在，通过词条的上位词、下位词、同义词、近义词和相关词，可以较为精确地定位一个概念。词条间关系的基本类型有 3 种，分别是等同关系（equivalence relationship）、层级关系（hierarchical relationship）和相关关系（associative relationship）。其中等同关系和相关关系本身是没有方向性的，只是有些知识组织系统在做词汇控制的时候，需要做正式/非正式的区别。而层级关系是有方向的。但是随着知识组织系统的发

展，尤其是知识组织系统由面向加工标引人员逐步扩展到面向计算机自动化应用，更为细致的关系类型区分越来越有必要。等同关系在非标引领域可能会包含同义、近义、反义甚至部分上下位的类型，而层级关系也有部分-整体，类属和概念实例等子类，相关关系则更为复杂。一些相关国际和国外标准已经注意到这个变化，并且有一些推荐性的细分关系类型供参考。

在关系类型的扩展设计上，有自顶向下和自底向上两种思路。在汉语科技词系统设计中，两种思路相互交织，既保证关系类型知识体系的相对完整，又充分考虑了被处理的信息资源的实际分布情况。首先在已有的信息资源中进行采样，然后由知识工程师对这些采样进行分析，提取出可能的关系类型，由于知识工程师自身基础条件的差异，以及不同时间阶段的认识差异，形成的关系类型全集可能有多种粒度，也会有交叉，这个基础类型集的规模可能会较大，达到几百种到近千种。然后在此基础上做统一的粒度控制和意义控制，同时从逻辑上考虑在抽样中没有出现或者没有总结出来的关系类型，从而保证其完备性，经过归纳整理，一般一个领域其关系类型大概在几十种左右。

通过研究，我们给出 20 种扩展关系作为推荐的基本类型，此外，还可能存在领域独有的关系类型。当然，这 20 种扩展关系可能在不同领域表现上有一定差异，推荐关系列表如表 1-1 所示。

表 1-1　汉语科技词系统的推荐基本扩展关系类型

序号	关系类型	关系类型（英文）	序号	关系类型	关系类型（英文）
1	借助/利用	Ultilize/Be ultilized by	11	原材料/成品	Material/Product
2	类属/子类	Class/Subclass	12	原材料/设备	Material/Equipment
3	全称/缩略	Full name/Abbrevation	13	设备/消耗品	Equipment/Cosumble
4	基本等同[a]	Equivalent	14	过程/产品	Process/Product
5	部件有/组成	Component/The whole	15	过程/工具	Process/Tool
6	成分有/集成	Composition/The integrated	16	相似[a]	Similar
7	替代/被替代	Replace/Be raplaced by	17	配合[a]	Combine
8	继承/被继承	Inherit/Be inherited by	18	反义[a]	Antonymous
9	因/果	Causality/Effect	19	主体/附件	Main body/Attachment
10	影响/受影响	Affect/Be affected by	20	概念/实例	Concept/Individual

注：a 对称性关系类型，即关系是没有方向性的，如对于词条 A 和 B 有对称性关系 R，即如果 A-R-B，则必有关系 B-R-A。

2.3 词条的属性

属性也是一种对词条的限定，其描述的模式和关系一样，也是主体-谓词-客体这样的三元组，只是具体角色不同，关系描述三元组是词条-关系-词条，属性描述三元组是词条-属性-属性值。两者的不同：首先，三元组中客体对主体的依赖程度不同，虽然在

关系描述中，两个词条可能在关系中处于不同语义地位，但是它们作为词条的地位是相同的。当主体词条被删除的时候，这条关系消失了，但是客体词条依然存在。但是属性则不是这样，属性值是依赖主体和谓词而存在，除非属性值本身也是一个词条或其他的属性三元组的客体，主体被删除了，客体也将消失。其次，关系三元组的客体必须也是词条，而属性三元组的客体则约束比较宽泛，可能是词条，也可能是短语，甚至是一个句子。属性关系类型的设计与关系类似，也需要将自底向上和自顶向下结合起来，也分为基本的属性类型和领域相关的属性类型，汉语科技词系统推荐的属性类型如表 1-2 所示。

表 1-2　汉语科技词系统的推荐基本属性类型

序号	属性类型	属性类型（英文）	序号	属性类型	属性类型（英文）
1	特点	Characteristic	9	其他要求	Other Constraint
2	优点	Advantage	10	极限情况	Limited situation
3	缺点	Disadvantage	11	起始时间	Starting time
4	现状	Present situation	12	终止时间	Termination time
5	前景	Prospect	13	用途	Use
6	困境	Difficult future situation	14	功能	Function
7	数值范围	Number range	15	目的	Objective
8	数量范围	Quantitative range	16	方法	Method

2.4 词条的多维分类

汉语科技词系统支持多维分类和复分类，这意味着一个词条可以用不同的分类法加以标识，在同一分类法下，可以有一个分类号，也可以有多个分类号。从目前的实践及资源情况看，汉语科技词系统至少应该包含.3 个分类体系。首先是《中国图书馆分类法》（简称中图法，CLC），它主要面向图书资料；还有在科技文献中占主导地位的期刊文献，当前国内主要的期刊都有作者标注的 CLC 分类号；此外，像国家科技图书文献中心（NSTL）在加工的时候，也对部分外文期刊文献进行了 CLC 分类标注。其次，随着专利在科技创新领域的作用越来越大，国家也越来越重视，而专利通常使用《国际专利分类法》（IPC）加以标识。这两个分类体系都是各领域通用的。最后，每一个领域都会有自己独特的分类体系，能够弥补通用分类体系的不足，我们称之为领域相关分类体系（DSC）。DSC 通常是一个较为粗糙的 2～4 层的层级结构，一般有几十个到几百个类目。汉语科技词系统中的词都可以用以上这些分类体系进行多维分类。

分类有定性和定量分类两种方式，其中 DSC 分类号的赋予是定性的，即知识工程师和专家根据自己的理解来为词系统中的每个核心词进行设置的。而 CLC 和 IPC 分类号的赋予则既可以是定性的，也可以是定量的。定量的分类信息赋予，是与收集到的期刊文献和专利数据集密切相关的，虽然由于抽样存在一定偏差，分类号的赋予也可能有一定

的错误，但是希望这种概率分布尽可能接近真实情况。具体的计算方法如公式 1、公式 2 所示。

$$f(c_i|t)=\sum_N(g(c_i)\times(\sum_M(h(m)\times f_1(t,m))) \quad \text{（公式 1）}$$

$$p(c_i|t)=f(c_i|t)/(\sum_I f(c_i|t)) \quad \text{（公式 2）}$$

其中 $f(c_i|t)$表示在一个特定的拥有 CLC 或 IPC 标识的语料库中词条 t 关于类别 c_i 的概率，N 是语料库中的文档数量，$g(c_i)$是一个二值布尔函数，即如果某个特定的文档的分类标识中有 c_i，其值就为 1，否则就为 0。一篇文档通常会分为 M 个部分，这里 M 和文档的类型有关，如一般的期刊可能分为标题、摘要、关键词和正文 4 部分，而一个专利可能分为专利名称、专利摘要、专利权利要求书和专利说明书 4 个部分。$h(m)$是一个关于第 m 部分的重要性的参数，$f_1(t,m)$是第 m 部分中词条 t 的出现频率。I 是特定分类法，如 CLC 或者 IPC 分类法中类目的总数，$p(c_i|t)$是词条 t 隶属于类别 c_i 的概率。

定量的分类信息主要是与特定的文献分类方法结合的，通过定量数据，能够较为容易地判断特定类型文献的分类信息，为计算机自动或者半自动处理提供基础数据支撑。但是由于当前词系统语料库资源建设还比较落后，因此当前的分类号都是定性给出的。

2.5 词条的定义

定义是为了更加清楚地描述和定位词条而设定的。最初定义获取的主要来源是词典、百科全书等各种工具书，也包括期刊文献和互联网文献。通过研究团队实践，发现两类重要的来源——专业名词术语和标准，在开始时被忽略了。在这两类文献中，不是所有的术语都有定义，也不是所有的定义都完备，但是通常这类定义都更具有权威性，可以直接引用或者作为基础定义再由专家补充。

词汇的定义形式多样，常见的有内涵定义、外延定义（包括列举定义）、情景定义、理论定义、实物定义、本义狭义定义、递归定义、约定定义、劝导性定义等。本书研究主要考虑内涵定义和外延定义。

定义可能有一条，也可能有多条，这是因为对于同一个词条，可能存在多条具有代表性的权威定义，这是科学的和客观的。如果定义有明确出处，则通常以“【来源】”来分隔定义和出处。通过阅读出处参考文献可以更全面地了解这条定义。

在定义处理上，可以采取一些辅助工具。首先要确定各种定义的模式和类型，由于定义相对规范，因此特征的提取并不困难，这一工作将有助于进一步自动化地抽取定义，从而能够节省专家的劳动，让专家将工作的重点放在定义准确性的审核上。

2.6 形式化概念描述

形式化概念描述是应对计算机环境下对词条较为重要的一项知识，有利于自动化的内容分析和计算。在如叙词表、本体等典型的知识组织系统中，有很多的关系描述，这

些关系描述能够清晰地定位概念，标引加工人员使用起来非常方便有效。但是目前信息技术的发展和海量信息资源的存在决定了越来越多的应用要依赖计算机自动完成或者辅助人工完成。在这些情况下，单纯依赖关系定位概念，使用时就不够方便，因此有必要针对计算机处理提供一些形式化的概念描述方式。在概念描述的过程中，我们主要是受到字、词典编纂过程的启发。通常编纂者会选出一个有限的词或者字的集合，即规模为一两千的常用的字词集，之后所有词条和字的解释都依赖于这个有限集合，字、词典编纂及使用的实践证明这种方式是有效的。但是一个较大的问题在于如果还是用自然语言的字词，计算机理解起来还有困难，而形式化的描述对计算机效果会更好。也就是说，我们需要一个有限的形式化符号集合来描述每个词条代表的概念。由于计算机处理能力的增强，这个基本符号集合在数量上可以超越词典编纂的词集。在目前中文概念形式化描述中，有 3 种较为先进的方案，分别是上海交通大学陆汝占教授提出的内涵逻辑的概念描述方法、中国科学院董振东教授提出的基于知网（HowNet）的基于义元的描述方式，以及中国科学院黄曾阳教授提出的基于概念层次网络理论（Hierarchical Network of Concepts，HNC）的基元符号描述体系。由于现在还不清楚第一种方法的基元数量，因此后面两种方法较为可行，这里我们采用基于 HNC 理论概念基元的方式来描述汉语科技词系统的词条。

汉语和大多数西方语言不同，大部分的汉字不仅是形符、音符还是义符。所以在传统汉语理论中，汉语具有“聚字成词，聚词成段，聚段成章”的特点，同时也比较容易做到“见字知词，见词知段，见段知章”。所以传统汉语研究特别注重对汉字的研究，形成训诂学。在现代汉语研究中，徐通锵教授也提出了较为有影响的“字本位”学说，董振东和黄曾阳两位教授各自学说的提出也是受汉字的特点影响。如“政府”一词中，“政”意味着治理国家等政治相关事务，而“府”表示办事地点。假设不知道“政府”的含义，通过其组成的字的含义，也可以大致推断出词的含义。汉字本来是一字一义的，但是由于不同地域的影响，以及语言的发展，一字多义和一义多字开始出现。HNC 理论提出对概念的基元以字母和数字重新加以标识，从而解决了上述问题。HNC 的概念基元符号体系不是一棵树，而是由多棵树组成的一个森林状结构。在 HNC 理论中共有 18 个范畴、101 个组和 456 棵树。当然也可以用虚拟节点的方式，将其转化为一棵树，各层次的概念基元节点数共 6 580 个，所有这些节点，无论是否是叶子节点，皆可以用来描述概念。

概念描述是以一种半自动的方式来进行的，这种半自动的描述方式也有一定的局限，即不适用那些通过音译方式引入汉语的外来词。概念描述的基础是已有通用 HNC 语言知识库，语言知识库包含两个部分，一部分是 3 000 个单字-HNC 符号对；另一部分是规模在 4 万左右的词条-HNC 符号对，分别称为单字词库和多字词库。描述工具界面如图 1-1 所示，如对“新能源汽车”这一词条进行概念描述，通过查找已有词库，会分别发现关于“新”“能源”“汽车”的描述，可以把对应的符号作为“新能源汽车”的内涵加以描述。在 HNC 理论中，一个完整的概念符号是由若干基元和组合符号共同组成的，考虑到组合符号用法的复杂性和计算机处理的难度，因此在汉语科技词系统的实际概念描述中，仅仅考虑各个基元，而不考虑基元之间的组合关系。虽然这样会造成一定的语义损失，但是极大地降低了大规模描述的可行性和计算的复杂性。