中国图书馆学报(双月刊)2005年第5期
ZHONGGUO TUSHUGUANXUEBAO

●侯汉清 薛春香

用于中文信息自动分类的《中图法》知识库的构建

摘 要 中文文献数据库中存在着大量的分类号与关键词(或主题词)对应的人工标引记录。通过对这些数据的加工整理,以《中图法》类目体系为主干,组织各学科领域的语词,从而构建出反映分类号与语词概念对应关系的《中图法》知识库,用以实现信息的自动标引和自动分类。构建《中图法》知识库面临着一些难题:异构数据的整合;原始数据中分类号与主题词或词串之间一对多、多对多关系的筛选;标引词串与知识库中的词串的相符性比较等。图2。参考文献8。

关键词 《中国图书馆分类法》 《中国分类主题词表》 知识库 知识组织系统 自动标引 自动分类

分类号 G254

ABSTRACT The authors discuss the creation of a knowledge base of *Chinese Library Classification*, which contains the mapping of classification numbers and terms (keywords or subject terms), to realize the automatic indexing and classification of information. The knowledge base is to be generated from existing bibliographical databases containing manual indexing fields. However, there are some diffculties to be overcome. 2 figs. 8 refs.

KEY WORDS *Chinese Library Classification. Classified Chinese Thesaurus.* Knowledge base. Knowledge organization system. Automatic indexing. Automatic classification.

CLASS NUMBER G254

1 前言

随着计算机、网络技术的迅速发展,为了实现网络信息处理的智能化、自动化和精细化,以词表、分类表、语言形式出现的各种情报检索语言加快了与标记语言、超文本技术和其他软件技术的融合,出现了新一代的情报检索语言与自然语言的结合体——知识组织系统。

知识组织系统是指各种对人类知识结构进行表达和有组织阐述的语义工具,主要包括分类法、主题词表、语义网络、主题图、知识本体等[1]。分类法和主题词表在信息资源的加工组织中发挥着重要的作用,而语义网络、主题图、知识本体则是针对第二代语义Web提出的知识组织系统。本文所讨论的《中国图书馆分类法》(以下简称《中图法》)知识库也是一种知识组织系统,或称为用于自动标引和分类的专家系统,它建立在《中图法》的基础上,通过机器统计归纳出众多人工标引记录中所凝结的标引经验,建立分类号、主题词、关键词之间的概念对应关系,从而实现对文献的自动标引和自动分类,进而实现概念检索。

2 《中图法》知识库构建的原理

分类检索语言、主题检索语言和自然语言是3种不同的情报语言系统,标识和组织方式各不相同,但在本质上是一样的,都是一种主题概念标识系统,分类号、主题词、关键词都可用来表示某一文献信息的主题概念。因此,这三者之间存在着隐含的概念对应关系,即兼容关系[2]。

国内大多数图书馆、情报机构和信息中心所拥有的文献数据库中存在着大量的人工标引记录,这些记录中包含分类标引和主题标引(主题词串或关键词串)双重数据。我们可以通过对这些标引数据的计算机处理,挖掘出分类号—主题词串—关键词串之间的概念对应关系,实现三者之间的兼容互换[3]。在此基础上,构建一个自动标引和自动分类用知识库,实现中文文献的自然语言标引、主题规范、自动分类及概念检索。

不管是分类检索语言还是主题检索语言,乃至任何知识组织系统,都使用了分类方法。而《中图法》是一个建立在知识分类基础上、可用于信息组织的概念语义网络,因此,我们选择《中图法》作为本知识库的主干体系[4]。

(1)《中图法》是我国自编的一部大型综合性图书分类法,可用于图书资料、音像资料和其他类型信息的分类标引和检索。它在国内有着最广泛的影响和最众多的用户,是早已被大家公认的"不是标准"的标准。

(2)《中图法》自首次出版以来,在30多年里经过图书情报领域和其他各专业领域专家多次修订改版,具有广泛的学科覆盖面、完善的知识组织结构,在等级体系的基础上加入了分面组配的功能,能够适应现代文献信息分类自动化的

需求。《中图法》已建成了用最详细的元数据格式——MARC 描述的《中图法》数据库，2000 年出版了电子版，而且正在向网络版发展。

(3)目前国内几大文献数据库的分类标引均以《中图法》为分类依据，选择《中图法》作为知识库的组织框架，可以直接利用这些现已达到数百万、上千万条的标引记录，从而免去类号转换的麻烦。

(4)《中图法》从 90 年代起，已经实现了与国内规模最大、用户最多的叙词表——《汉语主题词表》(以下简称《汉表》)的兼容互换，研制并出版了国内最大的分类主题一体化词表——《中国分类主题词表》(简称《中分表》)。近几年还完成了《中分表》电子版的开发，并在新版中大幅度地增加了人口词，加快了检索语言的自然语言化。这一切为分类检索语言、主题检索语言、自然语言三者在标引、检索中的互操作奠定了基础。

(5)《中图法》的网络信息分类组织的可行性得到了大多数专家的认同，它也正在采取分面化、增加自然语言接口、增加超文本链接等多种措施，以适应网络信息组织的发展需求。

总之，在构建中文文献自动标引和自动分类系统用知识库时，选择《中图法》作为知识库的主干，具有明显的优势。

3　《中图法》知识库与《中图法》体系的结构比较

《中图法》与其他的传统分类法一样，包括分类表(含附表)和类目索引两大部分。随着情报检索语言向分类主题一体化方向发展，《中图法》与《汉表》融为一体。1993 年，《中图法》编委会在《中图法》与《汉表》对应的基础上编制出版了分类主题一体化词表——《中分表》，从而使《中图法》的体系日臻完善。《中图法》体系如图 1 所示。

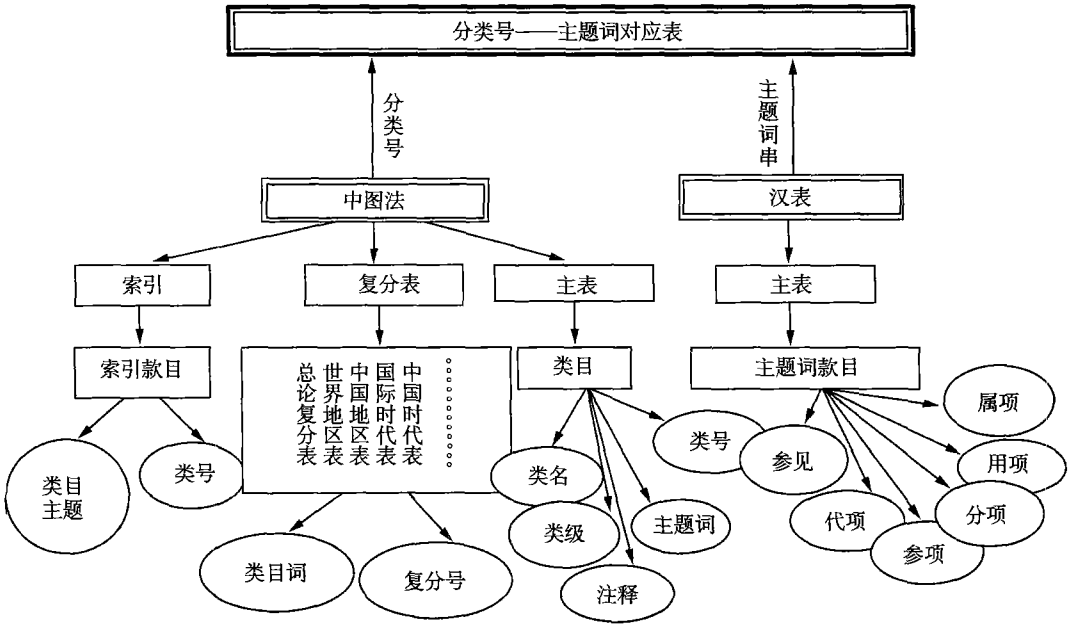

图 1　《中图法》的结构

《中图法》体系在文献手工标引时代做出了卓著的贡献，但在应用于网络信息和自动化时却暴露出如下弊端：

(1)无论是分类表还是对应的主题词表或分类主题一体化的《中分表》都属于受控语言，自然语言入口少，用户使用不便。

(2)《中图法》及其系统产品的着眼点是便于图书情报人员的标引和检索，而不是着眼于网络时代的普通用户的直接使用，因而过于强调词汇控制，忽略了检索语言与自然语言的结合。

(3)人工编制，定期修订，更新慢，大量新词、新主题、新类不能及时收录。

(4)受印刷版的限制，分类表、词表的规模偏小，类目和词汇数量少，难以满足计算机自动处理的需要。

由于现有《中图法》体系存在着上述弊端，所以有必要引进新的计算机和网络技术对《中图法》进行技术改造，包括采用计算语言学的方法和计算机编表技术，揭示分类检索语言、主题检索语言、自然语言之间的兼容互换关系，增加自然语言接口，扩充词表规模和加快增补更新，从而适应网络时代信息组织的发展。

我们开发的知识库以《中图法》为主干体系，包含若干

个词表和词典，其中分类号—关键词串对应表为主分类知识库，即知识库的主体；抽词词典、停用词表、同义词表、义类词典是主题标引知识库；地名表、时代表、文献类型表等为辅助分类知识库；另外还有采用《中图法》知识库标引和检索的文献数据库。图2显示了本知识库的结构。

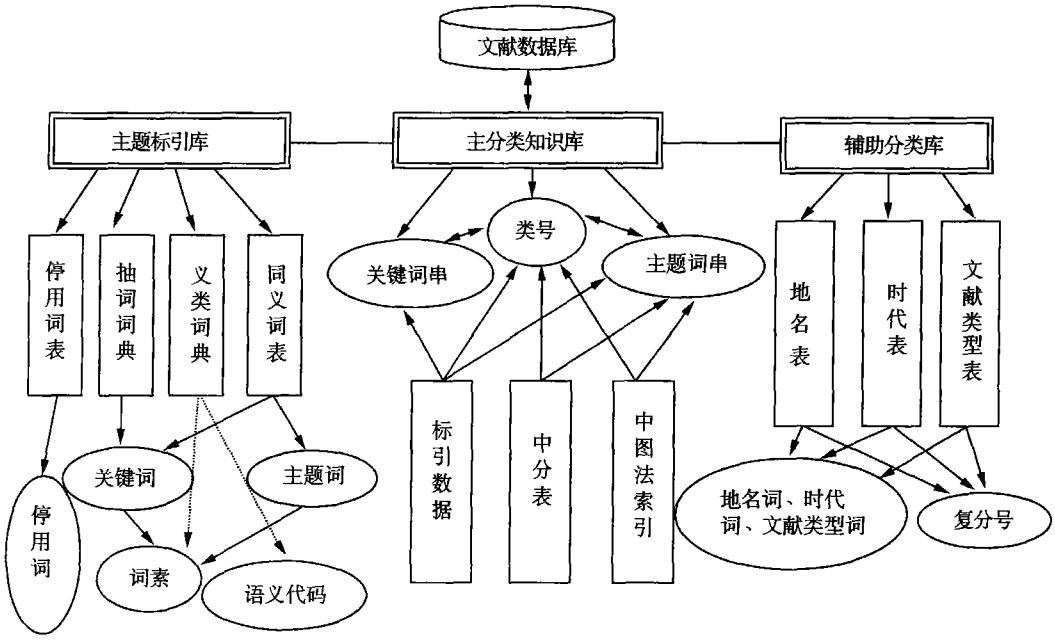

图2 《中图法》知识库的结构

图1、图2虽然都以《中图法》分类体系为主干，揭示主题词串与类号的对应关系，均可实现文献信息分类主题一体化标引。但两者相比，《中图法》知识库无论在内容、数量还是结构、功能上都优于《中图法》体系，更适合于文献信息标引的自动化和信息检索的智能化。

(1)《中图法》体系只揭示了分类号与主题词串的对应关系，而知识库则揭示了分类号与主题词串、分类号与关键词串、主题词与关键词之间的对应关系；后者词汇直接来源于文献，更新比前者快，便于用户检索。

(2)《中图法》体系只揭示了分类表中罗列出的类目和主题词串的对应关系；而知识库源于标引记录，包含了大量组合类目(仿分、复分时组配而成的类目)，类目数明显多于《中图法》体系。

(3)在《中图法》体系中一个类号最多对应20个主题词(串)，而《中图法》知识库中类目最多能对应几百个词串，平均一个类目对应10～14个词串，比《中分表》体系更能揭示类目的隐含概念；同时，词串数量大，便于自动分类中的相符性比较。

(4)《中图法》体系主要应用在手工标引和分类上；而知识库的规模大、容量大、更新快、可扩充性好，因而可以成功地应用于自动标引和自动分类。它不仅能保证较高的标引质量和分类正确性，而且在主题标引时不仅给出主题词还给出关键词，另外，它的同义词表、它的地名、时代、文献类型等的多重标引为信息检索提供了多个检索入口。

(5)传统的分类法和词表与书目信息都是分立的，而本知识库则与用其标引的文献相联通，即在某个类目体系下面直接就可以获取用该类号标引的文献信息，类似于网络环境下的目录型检索工具。

4 《中图法》知识库编制的关键技术

《中图法》知识库的研制面临以下几个难题：

(1)异构数据的整合。知识库原始数据主要来源于4类数据：原始类表数据，如《中图法》类目索引、《中分表》中分类号—主题词对应表；规范标引数据，即用《中图法》和《汉表》规范标引的书目数据，如上海图书馆的《全国报刊索引数据库》、国家图书馆、上海图书馆等的中文图书MARC数据；自由标引数据，即包含《中图法》类号和散标自由词的书目数据，如重庆维普的《中文科技期刊数据库》；题名库数据，从文献数据库的标引数据中取出题名和分类号构建而成。这4种数据描述的格式不同，有的是MARC，有的是文本，有的是数据表，词串之间的间隔符有空格、短横、冒号等，还有全半角之分。如何对这些数据进行整合，是构建原始库首先要解决的问题。

(2)一对多、多对多关系的筛选。原始数据中分类号与主题词或词串之间包含一对多，多对一和多对多的关系，而

本系统中必须设法为每一个词串确定一个惟一的分类号。

(3)标引词串与知识库中的词串的相符性比较。实际上二者完全匹配的几率是比较低的,所以本系统采用词汇相似度计算来实现概念标引、概念定类。如何从语义的角度来比较两个词或词串之间的相似度,而不是单纯从字面角度匹配,是实现知识库主题规范和自动分类亟需解决的难题。

针对上述难题,在编制和使用《中图法》知识库过程中应采用以下关键技术[5~6]:

第一,采用计算语言学的方法完成词表的构建。知识库原始数据主要来源于上述4类数据,首先要对这4类数据进行手工采集合并、删错去重,构建出原始库。原始库中包括类号与类名词、类号与主题词、类号与关键词的对应,从中分别抽取语词以及类号与语词的对应来构建知识库中的词表和词典。

在知识库中以分类号—关键词串对应表的构建最为关键,以计算语言学的方法来确定类号与词串之间的对应关系又是该对应表构造的关键技术。主要通过类目频次、词串频次、类号与词串共现频次的统计,采用数据挖掘中关联规则发现的两个参数——支持度和置信度来建立类号与词串的对应关系。

所谓支持度表示分类号和词串在整个原始库中同时出现的频度,即共现频次。共现频次越大,表示越多的标引员认可该分类号和词串之间的概念对应,那么这样的标引结果就可以认为具有普遍的正确性。

支持度 $Suppoort(keyword \Rightarrow clc) = P(clc, keyword) = freq\_gx$

其中:$P(clc, keyword)$表示在原始库中分类号和词串同时出现在一条记录中的概率;可用分类号和词串的共现频次 $freg\_gx$ 表示。

一般认为,支持度≥2表示该分类号与词串有概念上的对应关系,即有两人次以上认可这种对应关系。支持度越大,表示这两者之间概念对应关系成立的可能性越大。

而置信度则表示在出现该分类号的前提下出现该词串的概率。

置信度 $Conf(clc \Rightarrow keyword) = P(clc, keyword)/P(keyword) = Freq\_gx/freq\_keyword$

其中:$P(clc, keyword)$表示在原始库中分类号和词串同时出现在一条记录中的频度,即分类号和词串的共现频次 $freq\_gx$;

$P(keyword)$表示该词串在整个原始库中出现的概率,可用该词串在整个原始库中出现的频次 $freq\_keyword$ 表示。

同时满足最小支持度阈值和最小置信度阈值的规则称为强规则。当某一分类号和词串之间的支持度和置信度分别超过设定的阈值,则认为两者之间有很强的关联,即概念上的对应关系,以此来建立类号与词串的概念对应关系。

第二,通过相关度度量解决分类号与词串的多对一和多对多关系。在原始库中分类号与词串之间是一对多、多对一、多对多的关系,为给每一个词串确定一个惟一的分类号,需要度量分类号与词串之间的相关度。测量分类号与词串相关性的方法有多种,如信息对数量度法(IM)、极大似然法(LogL)、Dice测度等。我们基本采用Dice测度来计算词串对应的最佳类号。

$$Dice = \frac{P(clc, keyword)}{\frac{1}{2}[P(clc) + P(keyword)]} = 2 \times \frac{freq\_gx}{(freq\_clc + freq\_keyword)}$$

其中:$Dice$ 表示分类号与词串的并发概率,从而确定两者之间的关联度;

$P(clc)$表示该分类号在整个原始库中出现的概率,可用其在原始库中出现的频次 $freq\_clc$ 表示;

$P(keyword)$表示该词串在整个原始库中出现的概率,可用其在原始库中出现的频次 $freq\_keyword$ 表示;

$P(clc, keyword)$表示该分类号和词串在整个原始库同时出现的概率,可用其共现频次 $freq\_gx$ 表示。

在一个词串对应多个分类号的情况下,$Dice$ 值最大的记录表示该记录对应的分类号是该词串对应的最佳类号。

第三,构建义类词典进行词相似度的计算。主题标引从关键词转向正式主题词、自动分类中词串相似度匹配以及概念检索都离不开同义词的识别,因此需要在《同义词词林》[7]的基础上构造一个义类词典,通过语义编码从概念上识别同义词,而不是简单地通过字面相似识别同义词,是提高系统性能的关键之一。

《同义词词林》是一部按词汇语义分类的汉语词典,共14个大类、94个中类、1428个小类,以树型结构来表示词的语义关系。以它为基础,经过适当调整和编码,就可以构造出一部义类词典。《同义词词林》以单元词为主,其中大多可以作为构成复合词的词素。用它构建的义类词典一方面可以直接识别以单元词形式出现的同义词,另一方面以它作为语义工具,可以挖掘出以复合词形式出现的同义词和同义词组。

构造义类词典时,首先要将词汇的字面形式按其构成词素分解转换成语义代码,以《同义词词林》分类体系作为语义编码体系。

[语义编码]=〉(大类)(中类)(小类)(小组)

其中:大类=〉(大写英文字母);

中类=〉(大写英文字母)(小写英文字母);

小类=〉(大写英文字母)(小写英文字母)(数字)(数字);

小组=〉(大写英文字母)(小写英文字母)(数字)(数字)(数字)。

如:"商业"的语义编码为[Dil80203],其对应的大类、中类、小类、小组的编号分别为(D)、(Di)、(Dil802)、(Dil80203),其中"D"表示大类"抽象事物","Di"表示中类

"社会 政法",Dil802"表示小类 Dil8"事业 行业 工程"下的词群"行业","Dil80203"则表示小组"商业"。

有了义类词典,就可以对待识别的语词进行语义分析,把所有的词素归入相应的语义体系的结点之中,然后计算两个语词之间的语义距离,从而识别同义词和准同义词,实现从关键词向主题词的转换,并计算两个词串的相似度实现分类算法。

5 《中图法》知识库的使用

知识库以《中图法》为框架,以人工标引经验为基础,通过分类检索语言、主题检索语言、自然语言之间的兼容互换原理,建立分类号—主题词串—关键词串之间的对应关系,包含了丰富的词汇、大量的同义关系以及词串与类号的对应关系,能够广泛地应用于中文文献信息的自动标引、自动分类(归类),甚至概念检索上。目前,本系统已经比较成功地应用于网页和期刊论文的自动标引和自动分类[8],图书也在试验之中。

(1)利用抽词词典和停用词表进行分词,并借助于同义词表进行主题规范,实现中文信息的主题自动标引。

选择文献标引源,如题名、文摘、作者关键词、正文、参考文献等,利用抽词词典和停用词表采用最大正向匹配算法进行分词,统计词频、词数、位置权重进行排序输出标引词串,再结合同义词表进行主题规范,给出正式主题词。

(2)借助分类号—关键词串对应表、同义词表,以及地名表、时代表、文献类型表实现中文文献信息的自动分类。

本文说的自动分类是一种词串定类和概念定类,是一种基于实例的自动分类方法。首先,它是词串定类,而不是单词定类,提高了分类的正确性。其次,它是概念定类,在标引词串与分类知识库中词串进行匹配时,先进行字面相似度的计算,对于未能给出类号的记录再利用同义词表和义类词典进行语义相似度的计算,从而在兼顾正确性和速度的前提下,给出最佳的《中图法》主类号。第三,它是一种基于实例(即标引经验)的分类方法,分类知识库中的每一条记录都是一个标引实例,与其相匹配可确定其分类结果。第四,采用地名表、时代表、文献类型表对标引词串中的地名、时代、文献类型等分面归类,以改进《中图法》类目体系在自动分类上的弊端。

(3)在自动标引和自动分类结果的基础上,结合同义词表,实现中文文献信息的概念检索和多途径检索。

从标引的角度看,本系统给出的主题标引结果包括了关键词串和主题词串两个部分。用户一方面可以从关键词和主题词两个途径进行检索,另一方面能够实现词串检索而不仅仅是单个词的检索;此外还可以结合同义词表增加检索入口以及利用义类词典实现概念检索,从而提高检索效率。从分类角度看,分类结果包括了主类号以及地名、时代、文献类型等各个组面的复分号,用户可以从主题、地名、时代、文献类型等多个途径来进行文献信息的分类检索。

6 结语

《中图法》知识库是一个以《中图法》为主干而构建的知识组织系统,采用了中文文献数据库中存在的丰富的类号与词串的双重标引数据,具有良好的文献保障和用户保障基础。它将情报语言学的方法与计算语言学的方法结合起来,通过对大规模语料库的统计分析,利用计算机进行自动编制,克服了手工编制分类号—主题词对应表过程中产生的种种弊端。它基于《中图法》,却又比《中图法》具有更广泛的功能。它拥有丰富的词汇和语义关系,是一种基于概念语义网络的标引和检索用知识组织系统。

但是,它尚存在一些需要进一步解决的问题:

(1)知识库的完备性。《中图法》的固定类目是有限的,但是组配类目及其对应的词串则无法穷举。

(2)知识库的及时更新。包括及时添加新类、新词,未登录词的发现是一个亟待解决的问题;同时还要考虑陈旧类目和语词的淘汰问题,否则知识库过于臃肿会影响系统的性能。

(3)目前自动标引和分类使用的算法智能化程度仍不高,需要引入本体和主题图的一些技术以增加知识库的推理功能,改进知识库的性能。

(4)引入超链接、标记语言、互操作等技术,使知识库由静态走向动态,由线性走向网状,使知识库逐渐更新换代。

参考文献

1 曾蕾. 网络环境下的知识组织系统. 现代图书情报技术, 2004(1)
2 张琪玉. 关键词检索、概念检索与分类浏览检索一体化. 巨灵研究报告,2003-03
3 侯汉清,李波,戴晶萍. 计算机建立分类法和主题词表转换系统的尝试. 江苏高等学校图书馆学报,1999(4)
4 侯汉清. 建立以《中国分类主题词表》为核心的检索语言兼容体系. 见:21世纪高校图书馆的新使命——庆祝北京大学建校100周年国际研讨会论文集. 北京:北京大学出版社,1998
5 侯汉清,薛鹏军. 中文信息自动分类用知识库的设计与构建. 情报学报,2003,22(6)
6 章成志. 基于文本层次模型的Web概念挖掘研究——基于概念语义网络的自动标引和自动分类研究. 见:南京农业大学硕士毕业论文,2002
7 梅家驹等. 同义词词林. 上海:上海辞书出版社,1983
8 侯汉清,薛鹏军. 基于知识库的网页自动标引和自动分类系统. 大学图书馆学报,2004,22(1)

侯汉清 南京农业大学信息管理系教授,博士生导师。通信地址:南京。邮编210095。

薛春香 南京农业大学信息管理系博士研究生。通信地址同上。

(来稿时间:2004-12-01)