

● 李晓红 索传军

论电子图书馆的智能编目专家系统

ABSTRACT Intelligent cataloguing expert system is to integrate knowledge and experience of cataloguing experts into computer, process them and produce cataloguing results. It consists of the following five parts: knowledge database, integrated database, inference mechanism, human interface and knowledge acquisition. 2 figs 15 refs

KEY WORDS Electronic library. Intelligent cataloguing. Expert system.

CLASS NUMBER G254.36

随着计算机技术的发展,人们探求机器来模拟人类的智力活动成为可能,此种技术的实现即人工智能。人工智能(Artificial Intelligence)探求人类的思维过程,研究将人类的脑力劳动外延到某种物理装置,它是计算机科学的一个重要分支。而专家系统(Expert System)是人工智能走向实用化的一个最新研究领域。所谓专家系统,是利用存贮在计算机内的某一特定领域内人类专家的知识,来解决过去需要人类专家才能解决的实际问题的计算机系统。它是一种以知识为基础智能化的计算机软件系统。

1 智能编目专家系统的构成

智能编目专家系统(Intelligent Cataloging Expert System)是将编目专家的知识、经验加以总结,形成规则,存入计算机建立知识库,并采用一定的控制手段,对用户输入的信息进行处理,得出相应的编目结果,以供用户使用。一般说来,智能编目专家系统由五部分构成:知识库、综合数据库、推理机、人机接口和知识获取。

1.1 知识库

在对文献进行编目的过程中,由于需要数量巨大、种类繁多的知识,所以需将知识分别存于不同类型的知识库中,智能编目专家系统的知识主要分为以下五种类型:

(1) 综合型文献知识库:它包括有关文献的各种知识和信息。在信息时代,文献的内涵和外延都发生了巨大变化,它不仅包括一次文献(如图书、期刊等)、二次文献(如书目、索引、文摘等)、三次文献(如综述)和参考类文献(如百科全书、年鉴、手册等)等印刷型文献,而且包括电子函件、机读数据库等非印刷型文献,并且后者的数量在不断增长。综合型文献知识库还应包括文献之间的关系、文献元素(属性、句子、概念)之间的各种关系的描述。

(2) 专业知识库:它包括专业领域的概念及概念之间的各种关系。专业知识一般包含在各种专业词表(或词典)中,主要是主题词表。

(3) 规则库:它包括系统控制策略知识和编目专家的知识,可采用产生式规则表示法和框架表示法。规则库是系统执行推理解决问题的主要部分。

编目专家的知识主要有三类:第一类是编目常识,也就是编目专家的常识性的知识,

它来源于人类编目专家, 通过对编目专家编目实践的研究和分析, 从中发现规律, 对他们使用的常识性知识进行归纳, 输入计算机中。第二类是编目条例, 这是编目的依据。外文编目以 AACR2 (英美编目条例第二版) 作为著录规则, 中文编目以全国文献工作标准化技术委员会制定、国家标准局批准的有关各类型文献的著录规则为准; 第三类是编目特征识别知识, 通过这类知识, 可以判断各项编目特征 (如书名、著者、出版社等), 并由此决定编目规则的选择与利用。第三类知识是通过文献编目特征本身的研究获得的。在三级著录格式中, 目前一般都选用第一级著录格式, 其著录事项包括: 书名、责任者、出版地、出版者、出版时间、版本说明。

(4) 语言知识库: 用来存放编目过程中所需要的语言学方面的知识, 主要是字典、词典以及语法知识。语言知识库提供适合于任何专业领域的一切知识, 字词典知识一般可采用层次方法表示, 字词典库的每一个词条信息表示为一个层次结构, 一般有 3~4 层。词条信息分为若干款目, 表示意义的不同深度。最高层是词目层, 第二层是功能层, 第三层是意义层, 第四层是子意义层。只有具备好的字词典知识库, 才能支持提问和文献的自然语言处理和自然语言会话能力。

(5) 用户知识库: 它包括用户模型知识。用户知识库包括用户的专业知识、系统经验、常用的交互方式等, 其中专业知识是用户知识库和专业知识库的公共部分, 应使用相同的表示方法。用户知识库是辅助用户和提高系统效率的重要因素。

1.2 综合数据库

综合数据库用来存放编目过程中的大量事实数据以及关于问题求解的初始数据、求解状态、中间结果、假设、目标和最终求解结果。

1.3 推理机

应用各种编目策略和推理技术, 针对综合数据库中的当前信息, 识别和选取知识库

中对当前问题求解有用的知识进行推理, 以达到解决编目问题的目的。编目推理机可采取正向推理 (Forward Reasoning)、逆向推理 (Backward Reasoning) 或双向推理 (Bidirectional Reasoning) 方式。以上三种推理方式又分为精确推理 (Exact Reasoning) 和不精确推理 (Inexact Reasoning)。在编目专家系统中, 由于知识库中的知识往往是不完全的和精确的, 因而其推理过程一般采用不精确推理。

(1) 正向推理: 是指从可用的事实出发, 向前推理, 用当前的事实匹配规则的前提, 产生新的结论, 直到达到目标状态为止。这种推理方式是由数据到结论, 所以也称作数据驱动策略, 推理方式如下:

初始状态 目标状态
 (事实 条件) — (结论 假设)

(2) 逆向推理: 是指从目标出发向后推理, 用目标匹配规则的结论部分, 产生新的子目标, 然后试图证明一系列子目标, 直到找到支持结论的证据, 则停止推理。这种由结论到数据的策略也称作目标驱动策略。推理方式如下:

初始状态 目标状态
 (事实 条件) — (结论 假设)

(3) 双向推理: 指先根据原始数据, 通过正向推理帮助系统提出假设, 再运用逆向推理, 进一步寻找支持假设的证据, 如此反复这个过程, 直至达到目标。推理方式如下:

初始状态 目标状态
 (事实 条件) → ← (结论 假设)

(4) 精确推理: 在这种推理中, 领域知识具有确定性, 可表示成必然的因果关系和逻辑关系, 推理的结论或是肯定的或是否定的, 也可以把可能性大于某个固定值的假设认为是肯定的。

(5) 不精确推理: 在这种推理中, 使用确定和不确定的知识, 证据不一定是肯定的, 而是给予某种“权”(可信度)。推理的规则也

不是肯定的,也是给予某种“权”。对于多个证据或多条规则的推理要进行“权”的组合。由于不确定知识的表示方法及组合方法的不同,形成了各种不同的不精确推理方法。

1.4 人机接口

智能人机接口是最终用户、编目专家和知识工程师与信息知识源之间的中介,具有智能特性,能实现中间人专家所能完成的功能。

(1) 智能人机接口的特征:

第一,有智能性:接口能应用人工智能技术(如知识表示,语言理解,推理和学习等)模拟专家处理编目问题的认知功能,有效地执行各种认知活动,能从用户模型和编目模型推导信息,解释提问和补充问答,提供启发式策略,指导用户。

第二,有丰富的知识,如编目专家的知识 and 用户知识等。

第三,有较强的自然语言通讯和图形显示功能。

第四,对用户友好:允许用户以自然的方式表达需求,不受任何人工命令语言或语法的限制,并能处理不完整的、不精确的信息。

第五,适应性强:能适应不同任务的不同应用方面,也能适应于多种系统。它还适应于具有不同层次的、不同要求的各类用户,尤其是没有经验的新用户。

第六,能执行快速的对话,系统响应速度快。

(2) 智能接口的主要工作:

第一,问题描述:接受用户的问题及其他信息,进行一定的处理,产生需求模型或其他内部形式。操作处理应包括分析、理解问题,明白用户的需求,识别用户提供的各类概念,并进一步解释和表达问题。

第二,系统回答的表达和解释:将系统提出的问题,对用户问题的回答、推理结果及对结果的解释等进行逆转换,以用户可以理解的形式输出给用户。

第三,会话管理:会话管理问题可分为

两部分,第一部分是会话总体结构的控制,该部分描述会话的控制方式或执行计划,决定问什么问题,由谁(用户或系统)在何时提问或回答,第二部分是确定会话语言。

第四,知识获取:通过用户的交互对话,获取系统所需要的知识,其中用户知识即用户模型是最主要的。

1.5 知识获取

是指在智能编目专家系统的知识库建库中用以部分代替知识工程师进行专门知识的自动获取,实现专家系统的自我学习,不断完善知识库。获得知识的方法主要有两种。一是采访式。此法是由知识工程师通过一系列的采访和编码从人类编目专家那里获得知识,知识工程师是专家和计算机系统之间的中间人,他在知识获取处理中起核心作用,知识的质量很大程度上依赖于知识工程师的技能。在获取知识和建立知识库的过程中,知识工程师经常要面对面地与专家会话,会话允许专家综合性地讲述任务和他们的专知,这些综合性的讨论应事先计划好,以便有详细的目标和问题。由于专知的特殊性,还没有确定的方法引导会话,个人之间的通讯和分析技巧常被用来引导专家完满地回答询问,并合理地解释这些回答。二是机器学习。机器学习是知识系统利用各种学习方法来获取知识,并进行知识的积累,建立、修改和补充知识库。它是一种高级的全自动化的知识获取方式,减少了对专家的和知识工程师的依赖性。在机器学习中,机器归纳学习很受重视,并已取得了一些成功的经验。

2 主题词表和分类表的建立

2.1 主题词表的建立

(1) 主题词是能用于描述、存储、查找文献而作为建立检索工具依据的规范化的词汇,是主题词表中表达一定意义的、最小的词汇单元。主题词一般分为标题词、元词、关键词

和叙词。

主题词表是把主题词按照一定的方式排列而成的词汇表。主题词表(词汇)一般由详表(包括主表和副表)、简表、索引和词图等部分组成。

由于主题词表是按情报检索查全和查准的要求,通过明确概念(即主题)之间的相互关系的方式组织主题词的,所以它可以作为标引、检索和查询文献的依据,主题词表是不同的文献标引人员和读者之间思维的桥梁,是图书情报部门开展工作的重要工具。

(2) 主题词表的知识表示主要是由智能编目专家系统的结构而定的。既可以把主题词表作为整个知识库的一部分,主题词表和知识库采用一致的知识表示方式,也可以把主题词表作为独立的知识库,可采用框架结构表示方法建库,这是因为主题词表中的主题词之间存在纵向和横向组合的关系。

框架是一种知识结构化的表示方法,也是一种定型状态的数据结构,它的顶层是固定的,表示某个固定的概念、对象或事件,其下层由一些称为槽的结构组成。每一个槽可以有任意有限数目的侧面,每个侧面又可以有任意数目的值,而且侧面还可以是其它框架(即子框架)。框架的一般形式如下:

《框架名》

《槽名 1》《侧面名 11》(值 111, 值 112...)

《侧面名 12》(值 121, 值 122...)

...

《槽名 2》《侧面名 21》(值 211, 值 212...)

《侧面名 22》(值 221, 值 222...)

...

...

《槽名 n》《侧面名 n1》(值 n11, 值 n12...)

《侧面名 n2》(值 n21, 值 n22...)

...

相互关联框架连接起来组成框架系统,或称框架网络。不同的框架网络又可通过信息检索网络组成更大的系统,代表一块完整的知识。主题词表的知识表示应做到尽可能使用户便于描述、推理和进行其它的智能操

作,便于对知识库进行管理,可以处理各种模糊信息。

由于主题词表的知识表示可采用语义网络或框架结构表示方法(实际上框架结构也是一种语义网络),所以它可采用 C 语言或关系型数据库来进行建库。下面以《汉语主题词表》为例,将主题词分成两部分,用两个库分别存放。

主题词表的第 1 部分

项 目 名	类 型
序 号	N
主 题 词	C
汉 语 拼 音	C
拼音编写码	C
词 标 识	N
英 文 名	C
范 畴 号	C
词 频	N

关键词库存放主题词的第 2 部分

项 目 名	类 型
序 号	N
关系符代码	N
关 系 词	C

主题词库以汉语拼音排序,其中词标志的值为:族首词——1,关键词——2,非正式主题词——3,自由词(用于机器索引)——4。关系词库中的关系符代码分别是:Y(用)、D(代)、F(分)、S(属)、C(参)、Z(族)。两库以序号项相关联。

(3) 主题词表的管理。为了适应事物自身发展变化以及科学技术发展变化的需要,为了保持生命力,任何一种词表或类表都需要不断地进行修改、补充,使之不断得到完善。主题词表也不例外,也将会随着自然语言处理技术的发展和专家知识的进一步获取而不断得到补充和完善。词表的变化主要包括主题词的增、删、调整语义关系等。

第一,对主题词及其相互关系进行增加、删除和修改。在主题词表的使用过程中,凡具有较多文献数量而又不能通过现有主题词的

组配作为标识时,应补充新的主题词。凡与主题词表中的其它主题词概念重复或者长期使用频率过低或者没有使用价值的主题词,应该删除。凡取名、分类、词间关系等存在矛盾或错误时,应该修改。

第二,建立主题词使用情况档案。主要包括三种:第一标引频率(即使用频率)统计档案;第二上位词标引和近义词标引档案;第三标引频率分析档案,不断提高主题词表的科学性和实用性。

第三,具有学习功能。能利用机器学习技术,不断从标引和检索过程的反馈信息中获取新主题词以及主题之间的关系,也可以用专家的启发式知识从词表本身推导出新知识,不断精炼和扩充词表知识库。

第四,主题词表的完全重新构造。当主题词表不能满足很多新用户的需要时,或者在修改词表过程中产生了新的词分类的方法时,可能需要完全重新构造主题词表。

2.2 分类表的建立

(1) 分类表是一种用于对文献进行分类标引的词典,它把类目按一定的方式组织起来,并用标记符号代表各级类目和固定类目的先后次序。分类表习惯上又称为文献分类法或图书分类法。分类表由三个要素构成:类目、类号和注释,三要素按一定的体系结构和技术方法组织起来构成了分类表。分类表主要有以下功能:

第一,分类表,尤其是目前我国通用的分类表都是按等级体系进行排列的,以类号作为文献主题概念的标识,按学科门类集中文献,为从学科专业查找文献提供了有效途径。

第二,分类表用等级结构显示主题概念间的关系,按学科体系排列类目,直接体现了知识分类的要求,把文献强制性地按特定分类体系予以组织,具有较强的系统性,便于检索某一学科、某专业领域内的文献资料,可以鸟瞰学科或专业的全貌,达到查找文献,由此及彼,触类旁通的目的。

(2) 如同主题词表,分类表的知识表示也是由智能编目专家系统的结构决定的。可以把分类表作为整个知识库的一部分,采用和知识库一致的知识表示方法,也可以把分类表作为独立的知识库,采用框架结构表示法。因为分类表是以知识分类基础按学科专业排列类目的,整个分类表中的类目构成了一个关联网络。分类表的知识表示也必须做到尽量方便用户的智能操作,便于对知识库进行管理。分类表也可以用 C 语言或关系型数据库进行建库,库结构如下:

项目名	类 型
序 号	N
类 目	C
类 号	C
注 释	C

(3) 分类表的管理。随着科技的发展和文献资料数量的增长,分类法要不断地进行修订和更新,分类表的有效管理是一部分类法永葆生命力的重要方式。

第一,加强对分类表构建管理工作,及时修改类号、类目中的错误,对于只有 A 类目参见 B 类目,而没有 B 类目参见 A 类目的反参照,及时予以补充,健全分类表中的参照系统,加强对分类表用语标准化规范化的管理,提高编表的质量。

第二,通过统计标引与检索的频率,可以发现那些使用频率很低或者根本没有人使用过的类目,对这样的类目应予以合并或删除,对于经常出现而分类表中根本没有的主题概念,应增加为新的类目,从而使分类表能很好地适应用户的需求。

第三,等级体系分类表利于从学科专业查找文献,却不能很好地按主题集中文献,所以字顺主题索引应成为分类表的一个重要组成部分。通过计算机可以自动生成字顺主题索引以及族系表和轮排表,还可以随时输出各种不同的分类表及其修订情况,从而使得分类表及其索引的编制速度和修订质量得到提高。

第四, 具有学习功能。应能通过机器学习技术, 不断地从分类标引和检索中获取新的类目, 用专家的启发式知识从分类表中推导出新的知识, 使得分类表知识库不断地得到完善。

2.3 分类主题一体化

(1) 所谓分类主题一体化是指分类法和主题法相互利用、相互结合、相互渗透, 形成一个分类主题词表。即分类语言和主题语言在概念系统、参照系统、标引系统、索引系统等方面合二为一。

分类法和主题法是从两条不同的途径对文献资料进行检索的两种不同方法, 两者各有所长, 又各有所短, 不能相互取代。为了扬长避短, 长期以来人们一直在探讨如何使两者结合起来, 相互取长补短, 从而满足用户多途径检索的需要。检索语言的发展史已显示出了分类法与主题法日趋结合的发展趋势。60年代叙词表的迅速发展, 更使主题法和分类法向着相互结合的方向迈了一大步。叙词表不仅完善了参照体系, 而且增加了范畴表、族系表等部分, 在叙词表中运用了很多分类的方法。60年代杜定友先生提出编制分类主题目录的建议, 但没有得到国内同仁的重视。英国的艾奇逊(Aitchison, J)主编的《分面叙词表》(Thesaurofacet)第一次实现了分类主题的一体化。他的词表包括分类表和叙词表两部分, 每个叙词同时出现在分类表和叙词表中, 两部分之间用分类号相联系。此外美国、日本、德国等国家也在这方面进行了大量研究, 使得分类主题一体化成了当代图书情报界热门的研究课题之一。

(2) 虽然分类法和主题法在外部特征、组织方式、标引方法等方面均有不同, 但它们在原理上有诸多相同之处, 从而成为分类主题一体化的基础。

第一, 两者皆使用了分类的方法。分类表和主题词表的编制, 以及主题分析和文献标引, 都需要运用分类的方法, 只是主题法是通

过参照系统、范畴索引、词族索引等方式来运用分类方法的, 而不像分类法那么直接公开地采用分类的方法。

第二, 从发展历史看, 两者已从早期的互不相同发展到现在的相互渗透, 合二为一。

第三, 两者的最重要的构成要素分类号(或类目)和主题词在本质上是相同的, 都是代表一组相同主题文献的类集, 是文献主题概念的字面表示形式。

(3) 分类主题一体化的优点:

第一, 用户可以在一个检索系统中进行分类和主题两种方式的查询, 从而提高了检索效率, 节省了时间。

第二, 标引人员可以同时完成对文献的分类标引和主题标引, 两种标引结果可以相互转换, 节省了人力和物力。

第三, 检索语言的编制可由人工智能和计算机自动进行, 以提高编表的速度和质量。

第四, 检索语言的管理工作可由一个机构集中统一进行, 有利于词表的统一编制与修订, 便于管理和提高词表的质量。

(4) 分类主题词表的编制。分类主题词表主要由两部分构成: 分类号—主题词对应表和主题词—分类号对应表。若使用《中图法》, 可把《中图法》分类表改成分面分类表, 把主题词对应到相应的分面中, 形成分类号—主题词对应表, 这个表按分类顺序排列, 主要供检索使用。

主题词—分类号对应表是主题词与分类号的对照, 可由分类号—主题词对应表逆转过来编成, 按词的字顺排列, 用于对文献进行自动分类标引。

3 联机编目与查询

3.1 文献编目

所谓文献编目, 是指款目编制和目录组织管理的全过程, 款目编制即文献著录, 是对文献的物质形态进行分析、选择和描述, 对文

献的内容特征进行文献分类和主题标引的全过程,它分为描述性著录和检索点的选择两部分内容。著录的依据是编目条例及编目规则,通过著录形成款目,把款目按照一定的排检方式和组织原则结合起来,形成目录体系。按组织方式的不同,目录可分为两大类:分类目录和字顺目录(包括书名目录、著者目录、主题目录等)。

编目是图书馆的一项重要业务工作,编目的成果——目录是图书馆业务工作和读者服务工作的重要工具。图书文献的采购、编目、典藏、宣传报导等工作都离不开图书馆目录。目录更是读者从浩如烟海的文献中查找自己所需知识的主要手段。图书馆目录又是进行文献信息交流,实现资源共享的基础,尤其是在实现了图书馆自动化和网络化的前提下,更是如此。

3.2 联机编目与查询

(1) 联机编目是指借助计算机和现代化通讯技术所形成的联机书目设施,对图书文献进行款目编制和目录组织,各图书馆分担编目任务,共建中央数据库,达到任何一个图书馆编目,其他各图书馆共享编目成果的目的。

联机编目是计算机网络技术和图书馆业务工作相结合的产物。它有以下作用:

第一,联机编目汇集各个图书馆的藏书记录,形成中央数据库,提供联机查询和藏书地址等,可以起到资源共享的作用。

第二,联机编目可以使任何一个图书馆的编目记录被其他图书馆所利用,避免了各个图书馆分别编目的重复劳动,节省了人力、物力和财力。

第三,联机编目对任何一种图书文献编目一次,避免了以往对同种文献各个图书馆采取不同的著录级次,著录不同的项目的著录方式,有利于实现书目著录标准化,提高书目数据库的质量。

第四,联机编目推动了图书馆自动化的进程。图书馆自动化建设的前期工作是建立

书目数据库,这需要投入大量的人力、物力和财力,尤其对资金相对短缺的中小图书馆而言,更是难以承受。而一些有经济实力的图书馆首先建库,小图书馆可以直接上网利用中央数据库中的书目记录。

(2) 机读目录的形成。通过光符阅读器自动快速地阅读书名页、目次页甚至图书的全文,计算机自动地记录下此种图书的书名、著者、版次、出版地、出版者、出版时间、开本页码、关键词等书目的特征,利用联机查询系统,查询中央数据库中是否有与此种图书的书目信息相匹配的记录,如果有,则此种图书为复本图书,不予重新编目,自动地给予此书的馆藏地址即可。如果与中央数据库的记录不匹配,则此种书为新书,需要对此种图书进行编目。利用智能编目专家系统的知识库中的知识和规则自动地规范化地著录书名、著者、版本等描述性编目信息,根据所获得的关键词,通过分类表和主题词表数据库等进行自动分类和主题的自动标引。然后自动地在各个字前填写指示符、分隔符、记录终止符等,形成一条书目记录,自动存入中央书目数据库中,并能根据需要自动地进行屏幕输出、打印输出、磁盘磁带输出、光盘输出等形式的机读目录。

(3) 自动分类。分类是对文献进行标引的一种重要方式。传统的手工分类方式由于分类标准难以掌握,工作量很大,工作效率也非常低,并且所分类目不够细致全面,大多数文献只归入一类,难以全面揭示文献的多个主题,而自动分类却能解决这些问题,完善了文献的分类标引方式。国外自动分类的研究始于 50 年代末期,70 年代中期进入实用阶段。近年来我国也开始重视对中文文献进行自动分类的研究,并取得了很大进展,其中上海交通大学研制的自动分类系统,广东中山图书馆研制的我国第一个地区性公共图书馆自动化系统 ZSLA IS 中的自动分类系统已投入使用。

上海交大的自动分类系统主要有三个关

键部分: 抽取题中关键词、构造分类用主题词表、自动给定类号。对国内中文自然科学期刊的篇名和论文主题的匹配情况进行统计, 完全匹配的占 89.2%, 部分匹配的占 10.8%, 完全不匹配的仅占 0.23%。而中山图书馆的自动分类系统是在《中国分类主题词表》机读版的基础上进行二次开发研制而成的。先选出文献的关键词, 输入关键词后, 系统即以前方一致方式, 自动给出正式主题词和分类号, 从而实现了分类标引和主题标引同步完成, 大大地提高了工作效率和标准化程度, 保证了分类质量。

在智能编目专家系统中, 由于分别建立了分类表和主题词表作为知识库的组成部分, 所以可以从最能概括图书文献主题内容的题名中自动抽取关键词(中文抽词可以借助于用词词典和非用词词典, 通过用词词典抽出题名中有用的词, 通过非用词词典删除题名中无用的词), 与分类表的字顺主题索引进行匹配, 获得分类号, 完成自动分类的任务。

(4) 自动标引。自动标引是指计算机根据文献标题或文摘中的自然语言, 自动地选取表示文献内容特征的主题标识的一种文献揭示方式。与手工标引相比, 自动标引有以下优点: 第一, 标引速度快。日本的 JAKAS 系统早在 1982 年自动标引 200 万件题录, 共抽取 1000 万个关键词, 标引 1 万条题录只需要 0.35 小时, 而人工标引一篇文章大约需要 0.50 小时, 标引速度与计算机相差甚远。第二, 标引的一致性水平高。人工标引的一致性差, 标引结果常因人而异, 自动标引解决了这个难题。第三, 自动标引系统的检索效率已被证明优于或等于人工标引。由于自动标引有诸多好处, 国内外对自动标引的研究都比较重视。早在 1957 年美国 BM 公司的卢恩等人已率先撰文提出自动抽词标引的基本思想。目前西文文献的自动标引已经在实践中得到了普及应用。我国早在 1963 年就开始介绍国外自动标引的研究情况, 但直到 80 年

代, 汉语自动标引的研究才开始得到重视, 并成为我国情报学研究的重要组成部分。

自动标引的方法主要有两种类型: 抽词标引和赋词标引。所谓抽词标引, 是指用计算机从文献中自动地抽出能表达文献内容主题标识的一种自动标引。所谓赋词标引, 是指依据文献中的词汇特征, 从规范化的主题词表中选出相应的主题词赋于该文献的一种自动标引。而目前研究较多, 比较成熟的是抽词标引。抽词标引法又分为主关键词抽词标引和全关键词抽词标引。主关键词抽词标引是指计算机从题名、文摘甚至正文中抽取能表达文献内容的主要关键词, 作为文献的主题标识的一种抽词标引。全关键词抽词标引是指计算机是从文献中通过非用词表剔除非关键词, 余下的全部关键词都作为文献主题标识的抽词标引。

由于中文不同于西文, 词与词之间没有分隔符, 就使得中文文献自动标引比西文更加困难。自动分词是中文文献自动标引最困难的问题。关于中文自动分词问题, 近年来提出了不少方案, 主要分为两类: 基于算法的分词法和基于知识的分词法, 研究工作取得了很大进展。

总之, 在智能编目专家系统中, 由于已建立了作为知识库的一部分的主题词表数据库, 可以把读入计算机中的能表达图书文献的主题内容的关键词和主题词表按前方一致的方法进行匹配, 从而自动地把匹配的主题词抽出来, 完成自动标引任务。但如果建立了分类主题词表, 可以采取以上方式一次抽取主题词和分类号, 同时完成自动分类和主题标引两项工作, 提高工作效率和工作质量。

参考文献

- 1 Roy Meador III, Glenn R. Witting AACR2 Rules Used in Assigning Access Points for Books in Two Subjects: Implications for Automatic Cataloguing Expert (下转第 66 页)

4 结 论

将本文确定的图书馆学、情报学领域的核心期刊与南京大学《核心期刊目录》中的相比,可得出以下结论:

(1) 本文确定的核心期刊在数量上是 8 种,基本符合布拉德福定律 ($8 \div 213 \approx 1 \div 30$);而南京大学《核心期刊目录》确定的有关图书馆学、情报学领域的核心刊物是 14 种,在数量上过多 ($14 \div 213 \approx 1 \div 15 > 1 \div 30$)。

(2) 从表 1 中的数据可以看出,南京大学《核心期刊目录》确定的有关图书馆学、情

报学领域的 14 种核心期刊,其中有的无论在转载量或被引次数上都未达到核心期刊应有的标准。

(3) 从表 1 中的数据来看《中国图书馆学报》,无论其被转载量或被引次数都位居第一,是名副其实的一流刊物。

(4) 《图书情报工作》的被转载量和被引次数在表 1 中都位居第二,也应列为一流刊物。

夏明顺 南京大学历史系资料室工作。通讯地址:江苏南京,邮编:210093。

(来稿时间:1997.09.11。编发者:刘喜申)

(上接第 53 页)

- System. Library Resources & Technical Services, 1991, (2): 135
- 2 Roy Davies, Brian James Towards an Expert System for Cataloging: Some Experiments based on AACR2 Program. 1984, 8(18): 295
- 3 Susan Gauch, John B. Smith An Expert System for Automatic Query Refomation. Journal of the American Society for Infromation Science 1993, (3): 124
- 4 Micheline Beaulieu, Christine L. Borgman A New Era for OPAC Research: Introduction to Special Topic Issue on Current Research in On-line Public Access Systems Journal of the American Society for Information Science 1996, (7): 491
- 5 Carol A. Hert User Goals on an Online Public Access /catalog Journal of the Am rican Society for Information Science 1996, (7): 504
- 6 Ingrid Hsieh-Yee Student Use of Online Catalogs and Other Information Channels College & Research Libraries 1996, (2): 161
- 7 Margie Epple, Bernice Ginder Automated Systems and Subcolletion Designation Library Resources & Technical Services 1991, (2): 170
- 8 Songqiao Liu, Elaine Svenonius DORS: DDC Online Retrieval System. Library Resources & Technical Servicel 1991, (4): 359
- 9 Mark L. Grover Cooperative Cataloging of Latin-American Boods: The Unfulfilled Promise Library Resources & Technical Services 1991, (4): 406
- 10 张琪玉 分类法主题法一体化自动标引系统的基本原理和方法 图书馆学·信息科学·资料工作 1996, (3): 25
- 11 祁延莉 计算机辅助数据库标引技术的现状 图书与情报, 1997, (2): 34
- 12 莫少强, 许兴兴 计算机辅助图书分类、主题标引与检索系统 中国图书馆学报, 1995, 21(5): 41
- 13 张云芳 国外编目研究工作新进展浅探 图书馆论坛, 1995, (5): 41
- 14 马玉祥, 武波 专家系统 成都: 电子科技大学出版社, 1994
- 15 张玉峰编著 智能情报系统 武汉: 武汉大学出版社, 1991
- 李晓红 武汉大学图书情报学院博士生。通讯地址: 湖北武汉。邮编: 430072。
- 索传军 郑州工业大学图书馆馆员, 业务副馆长。通讯地址: 河南郑州。邮编: 450002。
- (来稿时间: 1997. 11. 4, 编发者: 赵薇)