基于自动标引的《中分表》改造及测评研究

何　琳　何　娟　阎素兰

摘　要　本文探讨了基于自动标引的《中国分类主题词表》(简称《中分表》)改造的模式、结构以及关键技术。在原《中分表》分类体系的框架之上,收集标引经验库中分类标引和主题标引的双重标引数据及其他相关数据,应用支持度、置信度和相关度等筛选处理方法,最终得出分类号与关键词(串)的最佳对应关系组合。本文从收词量、相符度、专指度、标引深度、主题标引能力和分类标引能力6个方面详细地对改造后的《中分表》进行了测试,结果表明改造后的《中分表》在编制方式、类目设置、收词量、全面性和专指性等方面都具有一定优势。建议在《中分表》的更新改造中,尽量采用立体化的整体结构,保证完备的收词量,进行必要的分级化控制并扩大用户交互。表7。参考文献9。

关键词　《中国分类主题词表》　自动标引　自动构建　测评指标

分类号　G254

Research on Revision and Evaluation of CCT Based on Automatic Indexing Data

He Lin, He Juan & Yan Sulan

ABSTRACT　This paper summarized research results on the automatic construction of CCT based on indexing data, which focus on construction mode, framework and main approaches. It collected all of the classification and subject indexing data from indexing database based on the framework of CCT to find the best corresponding of class number to descriptor by filtering processing method of confidence, support and degree of association. It also tested the construction results by vocabulary, association, specificity, indexing depth and external data in details. From the testing results, it proved that the revised CCT has some advantages on construction mode, class setting, vocabulary, comprehensive and specificity. At last, the paper suggested that it is very important for the construction of CCT to take a three-dimensional construction in achieving complete vocabulary, grade-based management and sufficient user interaction. 7 tabs. 9 refs.

KEY WORDS　CCT.　Automatic indexing.　Automatic construction.　Testing indicator.

《中国分类主题词表》(简称《中分表》)实现了文献信息的分类主题标引一体化,在图书馆、情报机构和信息中心等机构都得到了广泛的应用。然而随着时间的推移、科技不断发展、用户信息需求发生变化,《中分表》体系的弊端日益显露出来,如类目体系设置陈旧,修订更新周期漫长,篇幅长度受限,难以冲破编表专家的认知模式等。为了使《中分表》能够跟上时代的步伐,及时增补新条目,使条目更加贴近真实的文本特征,本文以原手工编制的《中分表》为框架,对其进行了基于自动标引的改造研究,并在此基础上对其性能进行详尽测评以发现改造研究的优势与不足,以期为《中分表》的修订提供一些建议。

1　改造的总体设计

1.1　构建原理

基于自动标引的《中分表》改造是指以情报检索语言中的分类语言、主题语言、自然语言三者之间的兼容互换原理为依据,采用计算机技术从标引经验库中获取所需标引记录,并运用

通讯作者:阎素兰,Email: yansl@njau.edu.cn

各种算法对所得数据进行加工处理,最终构建出能够反映分类号和关键词(串)之间对应关系的知识组织系统。

其中标引经验库即各种中文文献数据库,如中国知网、维普数据库、万方数据资源系统等,这些数据库都是专业图书情报人员对图书、期刊论文等进行著录、标引的数字资源集合,蕴含了丰富的人工标引经验。本文提出的基于自动标引的《中分表》改造就是在原《中分表》分类体系的框架基础之上,搜集标引经验库中存在的大量分类标引和主题标引的双重标引数据及其他相关数据,通过测算这些数据的支持度、置信度和相关度,对其进行筛选处理,最终得出分类号与关键词(串)的最佳对应关系组合。

1.2 构建模式

改造后的《中分表》是一个以通用分类体系《中国图书馆分类法》(简称《中图法》)为核心,实现各种分类语言、主题语言及自然语言之间的集成系统。这种构建模式应做到上下兼容,一方面向上可以为本体构建提供有力的基础数据,另一方面可以为各种专业词表的构建提供兼容框架。因此,需要构建一个由若干词表、分类表、规范文档、词典以及相关的自然语言词汇组成的兼容体系,初步构建的时候依然是建立《中分表》和《汉语主题词表》(简称《汉表》)对照表,其中挂接相关的规范文档、词典和自然语言词汇,与其他分类表和主题词表兼容后可以根据不同需要进行映射或链接。

1.3 结构设计

1.3.1 主体结构

《中分表》仍以《中图法》为主干体系,包含若干个词表和词典,其中分类号——关键词串对应表为分类知识库;主题词典、抽词词典、停用词表、同义词表、义类词典为主题标引知识库;地名表、时代表、文献类型表等为辅助表[1]。

1.3.2 预留结构

《中分表》是实现不同分类表和词表的核心兼容体系,因此预留了字顺兼容矩阵和分类兼容矩阵,保留了其兼容接口。字顺兼容矩阵以《汉表》的全部正式叙词为主干纵向展示,标明其相应的《中图法》分类号,并可把其他参与兼容的叙词横向展示,列出其他叙词表中等值兼容或近似兼容的一个或多个叙词。分类兼容矩阵即从分类的角度描述,结构类似。

2 主要步骤与关键技术

2.1 数据提炼

数据提炼的主要目的是对标引经验库中大量的标引数据进行过滤和修正,去除标引款目中的无关词,并对错误数据进行过滤。

2.1.1 标引数据噪声词的过滤

在众多的标引记录中,标引人员的背景和能力差异较大,相当一部分标引记录中的关键词对主题标引的作用不大,因此需要首先将标引数据中的噪声词过滤掉才能够保证构建《中分表》的基础数据质量。噪声词的去除是将词串打散,直接计算每个标引词同分类号的归属度,将归属度低的标引词去掉再重新组合为词串,以这种方式来提高基础数据的质量[2]。

例如,词串"空间交会|轨迹|初始条件|特性分析|试验|数值仿真",计算出各词的归属度:

空间交会 V526 0.7226;轨迹 V526 0.412;试验 V526 0.0052;数值仿真 V526 0.262;初始条件 V526 0.0025;特征分析 V526 0.0031

通过设定的阈值将与类号相关度较小的"试验"、"初始条件"以及"特征分析"等过滤,保留标引能力较大的词汇。

2.1.2 错误标引数据的剔除

错误标引的数据即为分类号和主题词串存在多对一或多对多的关系。具体做法是为分类号和主题词串找到最佳的匹配对象,采用统计机器学习方法,设置了支持度、置信度和 Dice 测度三个指标进行过滤。

2.2 新词的发现和识别

新词的识别是为类目寻找匹配的未登录词,因此新词识别包含未登录词的识别(形式)以及与类目匹配词的确定(主题匹配)。

(1)候选词的识别

候选新词的识别首先采用 N-Gram 无词典文本标引方法获取语料中的关键词汇,然后利用汉语的词/词组的构成特点,手工建立一系列筛选规则,对 N-Gram 文本标引方法获取的数据进行过滤处理[3]。

以航空航天领域为例,由于父串和子串在进行比较的时候,生成了诸如“北斗在”、“按编队卫星”、“表面形貌上”等许多不完整词汇,需要通过手工建立的筛选规则进行词性规范。候选词筛选识别出的仅仅是形式上正确的词或词组,例如“半导体集成电路”、“科技部”以及“倒计时”等词汇,可以在多个领域内出现,因此需要进一步确定是否是领域内核心词。

(2)新词的最终确定

筛选方法是将候选词集作为抽词词典对标引数据再进行自动抽词标引,利用 TF/IDF 值做特征筛选,剩余的候选词作为最终的领域新词。这也符合文献保障原则,即识别的新词必须是在绝大多数文献中出现过的核心术语。由标引数据的分类号来确定识别新词所属的类目。通过特征筛选,将会把诸如上例中的“半导体集成电路”、“科技部”等“通用”词汇从航空航天领域内过滤掉。

2.3 语义互操作

在经过数据提炼和新词识别之后,得到了无数据冗余、含有类目新词的分类号——主题词对照高频表,语义互操作将根据原《中分表》的基础数据建立每个类目的特征模型向量,采用聚类的方式得到每个类目的扩展词串条目,以达到丰富和扩展《中分表》的目的。

2.3.1 类目特征向量矩阵的构建

对于类目特征向量的构建,采用基于规则和机器学习相结合的构建方法,把原《中分表》中的词串在去除了部分通用因素(根据类目需要)之后作为强规则加入到特征向量中,另一部分采用自动分类中特征词抽取的方法获取类目的特征向量。

2.3.2 聚类方法

类目特征向量采用关键词的形式描述了该类目的基本特征,聚类的方法则是以该特征为模板从经过处理的分类号——主题词对照高频表中发现类目的扩展关键词串,采用 KNN 聚类方法聚类,K 值的选取决定了建成的《中分表》的规模。

2.3.3 语义相似度的计算

类目特征向量与标引经验词串的相似度计算采取了基于语义的相似度计算方法[4],经过几年的发展,通过模式匹配、词典释义等方法从期刊论文、网页、维基数据等资源中获取并积累了大量的同义词和准同义词[5],为相似度的计算奠定了良好的基础。例如,大量释义中存在着如下模式:

＜句首号＞(简称|简称为|英文简称|中文简称|又称|又称为|亦称|亦叫|亦作|又叫|也称|也称为|俗称|又译|又译作|全称为|全称是|英文缩写为){左引号|冒号}＜候选同义词集＞{右引号}＜句子结束符＞;

＜句首号＞{是|是英文|即}＜候选同义词＞(的简称|的全称|的对称|的缩写|的英文缩写)＜句子结束符＞

基于此可以从语料中抽取出大量的同义词对,以加强词串间相似度计算的准确度。

3 编制结果测试与比较

为了获得全面可靠的测试结果,本文分别从基本指标测评和应用测评两个角度对改造后的《中分表》性能进行了测试。以下将经过标引人员手工编制而成的《中分表》称为人工表,将经过自动化编制技术改造后的《中分表》称为自动表。

3.1 基本指标测评

基本指标测评主要是对人工表和自动表进行横向定量比较,旨在发现两表之间存在的横向差异,进而对自动表进行修正。

3.1.1 测试数据及测试方案

在对两表进行对比分析时,本研究随机选取社会科学大类中的 F82 货币、G25 图书馆学、图书馆事业以及自然科学大类中的 R72 儿科学和 V4 航空(宇宙航行)这四个类目作为测试用数据集,通过比较选取收词量、相符度、专指度

和标引深度这四个通用评价指标对自动表和人工表进行比较[6]。

3.1.2 测评指标

(1)收词量

主题词表的规模和词汇的完备程度主要用收词数量来衡量[7],是衡量词表在领域内描述能力的一个重要指标。作为一部综合性的词表,需要满足综合性文献单位标引和检索的需要。统计结果(见表1)表明,自动表四个大类的标引词总数都要比人工表中的标引词总数多出至少1倍左右,因此可见自动表的规模要明显大于人工表,并且在词汇的完备程度上,自动表也占有绝对优势。

表1 人工表与自动表收词量比较结果(单位:个)

	F	G	R	V
自动表	443	1146	3029	4724
人工表	172	657	71	908

(2)相符度

测量自动表和人工表的相符度时,借鉴信息检索领域中基于用户的信息检索系统评价指标——覆盖率和新颖率。

所谓覆盖率就是指以类号为基本单位,自动表和人工表中相符标引词所占的比例。利用覆盖率,可以衡量出自动表和人工表的相符程度,从一个侧面反映出自动表的正确性。假设自动表中标引词的总数为 T_1,人工表的标引词总数则用 T_2 来表示,自动表和人工表中相同的标引词总数为 $T_{1\&2}$,那么两表的覆盖率计算公式即为:覆盖率 = $T_{1\&2}/T_2$。

其中,两个标引词相符的条件是:①两词为同义词或准同义词;②内涵相同、外延不同的两个同位词;③包含关系的两个词。

新颖率是衡量自动表中与人工表不一致的词汇与其描述类号的相关度,利用该指标可以反映词表在新增词汇方面的能力,计算公式为:新颖率 = $T_R/(T_1 - T_{1\&2})$。

其中,T_R 为与人工表不同的标引词与该类目相关的词汇总数。词汇相关性的判断由专业标引人员进行衡量。

统计结果如表2所示,从覆盖率指标来看,总体数值偏低,人工表中有近一半的标引词条目没有在自动表中体现,这反映出自动表在规范性方面有待进一步加强;但同时我们也发现了人工表中存在的问题,没有在自动表中反映的另外50%的人工表条目中一部分是文献保障率偏低,一部分是标引词过长,标引词数目少,过于宽泛。以上两个方面都是《中分表》更新时需要重点考虑之处。从学科角度来看,社科大类的覆盖率为55%,高于自科大类的34%,这和学科特点有关,社会科学领域中名词术语相对固定和规范,而自然科学领域发展迅速,新增科技术语较多。

表2 人工表与自动表相符度比较结果

	F	G	R	V	平均值
覆盖率	0.65	0.44	0.44	0.23	44%
新颖率	0.64	0.72	0.86	0.79	75%

从新颖率指标来看,自动表与人工表不相符的词汇中,近75%是与该类目相关,也就反映出自动表在反映新事物、新概念方面具有良好的表现,其中自科大类的新颖率82.5%高于社科类的68%,这也与覆盖率结果互为佐证。基于以上数据,建议在《中分表》的编制更新中,重点加强对科技名词术语的新词识别和发现及其词间关系的构建。表3为人工表和自动表V421.1类目对应标引款目对照。

表3 自动表增补标引条目对照表

类目	人工表	自动表
V421.1	火箭\系统结构\设计 火箭弹头 运载火箭构型	火箭弹头 B平面参数\探月轨道 飞行器外形设计\UG/KF\知识工程\特征造型 固体推进剂火箭\优化设计 飞行器不确定性设计\多学科设计优化\软设计理论 多学科设计优化\自动微分方法\灵敏度分析 运载火箭构型 可靠性设计\伺服系统\载人航天火箭\系统结构\设计 非数值型综合设计变量\飞行器

(3)专指度

专指度是指赋予文献的检索标识与文献实际论述主题概念的相符程度,本文用平均词长和先组度这两个指标来衡量词表的专指度。平均词长就是标引词的平均长度,计算公式为:

平均词长 = 标引词的总字数/标引词的总数

先组度的衡量采用词表中词串型的标引记录所占的比例来计算:

先组度 = 主题词串型标引记录的条数/标引记录的总条数

统计结果(见表4)表明,自动表平均词长小于人工表,且常用先组度高的词串型记录。在海量数据的今天,利用《中分表》进行文献的自动处理已经是大势所趋,自动表这种采用多个较短的标引词来标引类目的方式更适合计算机自动处理,这种方式无疑增加了匹配的入口,从而大大提高计算机自动匹配的结果;而人工表倾向于采用更为专指的长词来表示类目的主题,虽然标引结果经过专家的加工更加专指凝练,但从文献保障的角度来讲,和真实文本环境存在一定差距,需要增加匹配入口。

表4 人工表与自动表专指度比较结果

		人工表	自动表
平均词长	F	3.98	3.58
	G	3.93	3.61
	R	3.58	4.67
	V	4.63	4.55
	平均值	4.28	4.10
先组度	F	0.12	0.89
	G	0.22	0.95
	R	0.73	1.00
	V	0.46	0.99
	平均值	0.38	0.96

(4)标引深度

标引深度一般是指对文献内容进行周详标引的程度,简单地说,就是指标引一种文献所用的标识数量。就《中分表》而言,类目的标引深度可以体现出用《汉表》的主题词标引《中图法》的类目所包含的主题内容的周详程度,具体可以用两个指标来衡量:单条平均标引深度和类平均标引深度。单条平均标引深度就是指平均每条条目下包含的检索词总数,而类平均标引深度的计算公式为:类平均标引深度 = 主题词(串)总数/类目总数。

单条标引深度数值高表明每条款目拥有更多的标引词实现更为专指的描述,而类平均标引深度数值高则表明该大类拥有更多的标引条目来全面、详尽地表达该大类的内涵。统计结果表明(见表5),在自动表中,各大类的单条平均标引深度与类平均标引深度值都要大于其在人工表中的数值。这主要是因为,人工表采用人工标引的方式,成本相对于自动标引的机器成本要高很多,且不方便实现,费时费力,所以造成了人工标引的结果数量要少于自动标引。

表5 人工表与自动表标引深度比较结果

	单条平均标引深度				类平均标引深度			
	F	G	R	V	F	G	R	V
自动表	2.97	2.59	2.97	4.24	32.77	66.15	32.77	10.54
人工表	1.12	1.25	1.12	1.68	9.35	4.86	9.35	3.45

3.2 应用测评

应用测评主要是针对自动表本身进行纵向的评价。《中分表》编制的主要目的之一是方便快捷地实现分类标引和主题标引以及二者之间的兼容互换,基于此应用测评的目的是随机选取实际的标引数据,利用人工表和自动表分别进行分类和主题两种标引,以评价自动表在文献标引实践中的能力。

(1)主题标引能力测评

信息资源的标引是《中分表》的核心功能,为了反映自动表的主题标引能力,采用了试标引的方式,选取150条财政类期刊论文标引数据以及从互联网中下载的150篇财政类网页作为测试数据。利用主题标引评价指标,采用人工受控标引和基于自动表的自动标引对测评数据的标引结果进行了定量测评[8],依据标引性能对各项指标排名进行人工打分,性能指标最高

的为3分,排名第二的为2分,最低的为1分,得到的各性能指标得分情况见表6。

表6 受控标引方式和基于自动表的自动标引方式性能指标比较

	受控标引	自动标引
标引深度	2	3
标引一致性	2	3
标引专指度	2	1
标引速度	1	3
标引成本	1	3
标引员智力负担	1	3
语词更新速度	1	2
总分	10	18

在主题标引方面,无论对期刊论文数据还是网页数据,从各个指标来看基于自动表进行文本主题标引,都收到了良好的效果。

(2)分类标引能力测评

分类标引能力的测评主要是评价《中分表》对期刊论文、图书、网页等电子资源自动分类结果的正确率和召回率[9]。本研究随机选取《中分表》中所有大类涵盖的期刊论文各300篇进行自动分类测试,分类结果同人工标引分类号进行比较,测试结果见表7。从测试结果来看,基于《中分表》对文本进行自动分类的平均正确率在66%左右,对于类目庞大、专深的类目体系来讲,这种分类结果还是非常乐观的,如果将分类结果的类目进行深入细化,正确率还会有较大提升。

表7 《中分表》自动分类标引能力分布表

	正确率	召回率
社会科学部类	63%	98%
自然科学部类	68%	98%
综合	66%	98%

从结果也可以看出自动表存在的弱势,由于社科中的许多大类,如政治、经济、文化等类别,其类目中的特征词在其他的类别中也会出现,不能唯一地刻画类目特性,使得社科中许多类别相似类目的分类正确率大大降低。因此,从自动分类的实践中,不难发现自动构建的《中分表》在社科大类中还存在着类目的特征词语义描述不够精准、过于粗糙的问题,使得社科相似类目之间的区分度降低。

4 《中分表》编制及更新的几点讨论

4.1 改造后《中分表》的优势

通过上述对自动表的定量和定性分析可以看出,自动表较人工表在编制方式、类目设置、收词量、全面性和专指性等方面都具有一定优势,具体来说有以下两个方面:

4.1.1 词汇的收录更加完备

根据测评结果显示,自动表中四个大类的标引词总数都分别大于人工表的标引词总数,且升幅的最小值约为人工表的1.7倍,最大值则达到了人工表的42.7倍左右。这主要是因为自动表采用自动化编制技术,不但节省了大量时间,同时还可以更加方便快速地在各大中文文献数据库中搜集主题词,主题词的来源更广,且对于新词的适应能力也更强,从而就大大提高了词表的规模,增强了词表的完备性。

4.1.2 文献信息的表达更加高效

自动表无论是在专指度还是在标引深度等方面都优于人工表,表明了自动表更能提供与文献主题概念相符的检索标识;同时,人工表的单条平均标引深度和类平均标引深度都要低于自动表,说明自动表比人工表的分类体系更好地揭示类目的隐含概念,更能周详地表达文献信息的内容特征,提高标引的质量和效率。

4.2 对《中分表》更新的建议

自动表除了具有上述两点主要优势以外,还具有更新周期短、类目便于扩充、更有利于文本的自动分类和自动标引,以及更适用于一般用户等优点。但在新表中还存在着一些缺憾:比如,新表标引词的选取源于网络,且大量采用自然语言,没有经过一定的规范控制,会造成所

选标引词与类目之间没有直接对应关系的现象出现；新表的标引深度虽然大大高于人工表，但过高的标引深度也有可能会导致词汇的冗余等。下面就综合上述内容，针对《中分表》的修订与更新，提出四个方面的建议：

4.2.1 整体结构方面

传统《中分表》采用的是树状结构，从总到分、层层递进，有着很强的系统性，但也会造成同一个主题词被平行划分到多个不同的树状结构内，为标引和检索工作带来不便。因此，在修订过程中，应使《中分表》以立体方式展现，将某一领域内的知识元素按其内在的关联属性，以可视化和超链接技术揭示知识结构及其语义关系，使《中分表》中的概念体系呈现出立体网状结构，从而更好地适应网络信息资源迅速增多的现状。

4.2.2 词汇构成方面

在对《中分表》进行修订时，应尽可能保证词汇收集的全面性，不要刻意地控制标引词的数量，而是尽可能找全所有与类目相关的检索词，同时也不要对正式叙词和非正式叙词进行严格区分，要尽量穷尽所有同义词、近义词；对于标引词的来源，主要可以来自于用户在长期检索过程中积累下来的检索日志，以及各大中文文献数据库中选取出的合适的标引词，以充分体现用户保障原则和文献保障原则；并且通过词频统计和聚类等技术对众多的词汇进行分门别类，即将原本一个整张大型词表拆分成由多个小型词表组成的知识组织体系，如基础性词汇组成基础词表、核心性词汇组成核心词表、专业性词汇组成专业词表等，这样一来就更加方便管理与更新，以及查找错误所在等。

4.2.3 描述格式方面

为了使修订后的词表能够用于自动标引、智能推理与检索等用途，应广泛使用 RDF、SKOS 或者 OWL 机器语言表达词表的概念关系，使《中分表》从最传统的纸质版本向由分类、主题和概念等不同语义级别的一系列词汇和概念数据库构成的集成知识体系转变，方便机器识别。

4.2.4 用户交互方面

通过构造一个用户参与的在线词表编制平台，使用户参与到词表的修订与更新中，同时配备专业的平台工作人员，对平台进行维护的同时，最重要的是要对用户提出的词汇进行审核，包括词汇的完整性、准确性和政治性等，再确定最终能够被纳入词表使用的语词，体现出用户的需求，发挥用户的积极作用，并且实现词表的动态更新。

5 结语

《中分表》体系规模庞大，在文献信息的分类主题一体化标引方面具有较强的性能，对我国图书情报界产生了广泛而深远的影响，但是随着科技与社会的向前发展，该表的不足和缺点正在日益显露出来，亟待采取一定措施弥补其中的漏洞。本文中所讨论的基于《中分表》的改造技术编制而成的自动表就是一个以人工表为主干构建出来的用于自动分类和自动标引的知识组织系统。它搜集众多中文文献数据库中存在的丰富的类号与关键词(串)的双重标引数据，具有良好的文献保障和用户保障基础。它将情报语言学和计算机语言学的方法结合起来，通过对大规模语料库的统计分析，利用计算机进行标识，克服了手工编制分类号与主题词对应过程中会产生的弊端。通过对其性能进行的一系列测评，可以发现自动表虽然基于人工表，却比人工表具有更广泛的功能。但在测评中也发现了自动表的一些缺陷，如标引词未经控制、易出现冗余等，这些问题还有待进一步探讨和解决。

参考文献：

[1] 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[J]. 中国图书馆学报，2005(5).
(Hou Hanqing，Xue Chunxiang. Creation of CLC knowledge base for automatic classification of chinese information

[J]. Journal of Library Science in China, 2005(5).)

[2] 何琳, 白振田, 侯汉清. 基于标引经验和机器学习相结合的多层自动分类[J]. 情报学报, 2006(12). (He Lin, Bai Zhentian, Hou Hanqing. Automatic multi-layer classification method based on integration of machine learning and indexing experience[J]. Journal of the China Society for Scientific and Technical Information, 2006(12).)

[3] 何琳. 领域本体的半自动构建及检索研究[M]. 南京: 东南大学出版社, 2009: 120 - 128. (He Lin. Research on semi-automatic construction and retrieval of domain ontology[M]. Nanjing: Southeast University Press, 2009: 120 - 128.)

[4] 侯汉清, 薛鹏军. 中文信息自动分类用知识库的设计与构建[J]. 情报学报, 2003(6). (Hou Hanqing, Xue Pengjun. Design & construction of knowledge database for automatic classification in chinese[J]. Journal of the China Society for Scientific and Technical Information, 2003(6).)

[5] 陆勇, 章成志, 侯汉清. 基于百科资源的多策略中文同义词自动抽取研究[J]. 中国图书馆学报, 2009(6). (Lu Yong, Zhang Chengzhi, Hou Hanqing. Using multiple hybrid strategies to extract chinese synonyms from encyclopedia resources[J]. Journal of Library Science in China, 2009(6).)

[6] 马张华, 侯汉清. 文献分类法主题法导论[M]. 北京: 北京图书馆出版社, 1999: 256. (Ma Zhanghua, Hou Hanqing. Introduction to library classification and subject indexing[M]. Beijing: Beijing Library Press, 1999: 256.)

[7] 侯汉清, 李华. 《中国分类主题词表》(第二版)评介[J]. 国家图书馆学刊, 2006(2). (Hou Hanqing, Li Hua. Evaluation and introduction of CCT(2nd)[J]. Journal of the National Library of China, 2006(2).)

[8] 刘竟, 朱书梅, 侯汉清. 网络环境信息标引的测评与比较研究[J]. 中国图书馆学报, 2008(1). (Liu Jing, Zhu Shumei, Hou Hanqing. A study of evaluation and comparison of information indexing in the networked environment[J]. Journal of Library Science in China, 2008(1).)

[9] 何琳, 刘竟, 侯汉清. 基于《中图法》的自动分类影响因素分析[J]. 中国图书馆学报, 2009(6). (He Lin, Liu Jing, Hou Hanqing. An analysis of the impact factors in the multi-layer automatic classification based on CLC[J]. Journal of Library Science in China, 2009(6).)

何 琳 南京农业大学信息管理系副教授。
通讯地址: 南京市卫岗1号南京农业大学信息科技学院。邮编: 210095。
何 娟 南京农业大学信息管理系硕士研究生。通讯地址同上。
阎素兰 南京农业大学信息管理系讲师。通讯地址同上。

(收稿日期: 2012 - 05 - 21)