●顾燕萍　侯汉清　王晓红

中文图书自动标引与分类加权设计研究*

摘　要　使用基于《中图法》知识库的中文信息自动标引和自动分类系统,对中文图书进行自动标引与自动分类的实验,以测试该系统对图书的适用性。实验通过对中文图书进行计算机自动标引与自动分类、人工打分测评、测试结果统计分析,得出中文图书的各标引源主题表达能力依次为:书名、内容提要、两级目次、参考文献、一级目次,在此基础上对标引源进行加权设计,权值设为5:3:2:2。实验证明该系统用于中文图书的自动标引与自动分类是可行的。表6。参考文献9。

关键词　中文图书　自动标引　自动分类　加权
分类号　G254

ABSTRACT　By using an automatic indexing and automatic classification system based on *Chinese Library Classification* knowledge base, the authors make an experiment of automatic indexing and automatic classification of Chinese books to test the suitability of the system for monographs. Through automatic indexing and classification of Chinese books, the results show the order of the subject representation degrees of various indexing sources: title, summary, two-level TOC, references, one-level TOC. Then, they get the weights of indexing sources 5:3:2:2. They think that the system is suitable for the automatic indexing and classification of Chinese books. 6 tabs. 9 refs.

KEY WORDS　Chinese books.　Automatic indexing.　Automatic classification.　Weighting.
CLASS NUMBER　G254

目前,中文文献标引大多仍采用手工标引和计算机辅助标引。面对巨大的文献出版量,手工标引逐渐暴露其缺点,严重影响我国文献数据库和网络信息检索工具建设的进展。因此,国内一些学者正在开发自动标引和自动分类系统。新闻出版署等单位研究的ECIP(电子在版编目)在电子书的基础上已有进展,但是电子书中缺少的主题词和分类号自动标引问题仍未解决。南京农业大学信息管理系近年来开发基于《中图法》知识库的中文信息自动标引和自动分类系统,此前已进行了报纸、网页、期刊论文等方面的应用研究[1-6]。本文利用上述系统,进行了图书的自动标引和自动分类的实验,以测试该系统对图书的适用性,并确定图书自动标引的标引源和标引加权方案,以便进一步改进。

1　基于《中图法》知识库的自动标引和自动分类系统介绍

该系统是一个基于《中图法》知识库的自动标引和自动分类系统,其中知识库是一个基于《中图法》的知识组织系统或者说是基于标引经验的知识库,包括了《中图法》库、《汉语主题词表》库、分类号—主题词(和关键词)对照库、同义词库等数据库。在此知识库的基础上,可以实现分类语言、主题语言和自然语言标引和检索的一体化,实现自动标引和自动分类。

该系统的技术路线是在确定基本信息标引源的基础上,运用基于词频的统计加权法,通过与分类号—主题词对照库主题词串的词面相似度计算,进而完成中文信息的分类标引。具体实施步骤:(1)提取作为标引源的文本信息,并放入数据库的不同字段备用。(2)词切分,即利用停用词库或半停用词库,将长字符串分割成若干较短的子串。(3)抽取关键词。抽词按正向最大匹配法进行分词,保证词长较大、专指的词汇能够优先抽出。(4)确定主题词,即将标引词由关键词转化为主题词,同时进行标引词的词频权值统计、排序,完成主题标引。(5)在分类号—主题词(或关键词)对照库的作用下,采用词面相似度算法,将主题词(或关键词)词串转化为相应的分类号,

* 本研究得到南京农业大学SRT计划基金(0413B02)和国家社会科学基金项目(05BTQ021)的资助。项目组成员还包括申卫国、兰锦生、夏海明。

完成分类标引[7]。

2 测试方案设计

2.1 测试的总体设计及其目标

先分析中文图书选取标引源,并确定经济类图书为测试对象,同时收集所需的各项数据;再拟定各标引源的权值及测试方案,然后利用系统对它们进行自动标引和自动分类,最后对测试结果作统计分析。本次测试的目标是通过对中文图书进行计算机自动标引与自动分类、人工打分测评、测试结果统计分析,得出中文图书的主题与图书的书名、内容提要、目次、参考文献等标引源之间的关系,分析不同标引源的主题表达能力,并在此基础上设计用于图书自动标引的相应权值。

2.2 标引源的选取及数据的收集

图书不同于报纸、期刊论文,篇幅通常比较长,至少百页以上。为提高运行速度,不宜将图书正文选作标引源。除正文外,图书一般还包括书名、目次、内容提要、前言(序言)、后记、参考文献等。经过调查分析,选取比较有代表性的几项(书名、内容提要、目次、参考文献)作为标引源。通过随机抽取,获得500种各项数据完备的经济类图书,将所需的各项数据转入数据库中相应的字段,建成一个access数据库以备测试使用。

2.3 测评和分级

(1)人工打分测评是将原来各标引源包含的关键词,与自动标引图书标引源给出的标引词,二者进行比较。由人工评判,给各个标引源的主题表达能力打分。打分的规则是按各标引源主题表达能力的强弱分为4个等级。一级:能很好反映图书的主题,给分为4;二级:基本能反映图书的主题,给分为3;三级:只能反映图书的局部主题,给分为2;四级:不能反映图书的主题,给分为1。

由于图书数据收集不完整,为了测试正确,只对500条数据中415种图书的完整数据人工打分测评。

(2)根据图书内容提要字数的多少,可分为4个等级。A:150字以内;B:150~250字之间;C:250~500字之间;D:500字以上。

目次按其规模可分为两级:一级目次包括部分(篇)、章,两级目次包括篇(部分)、章、节。本次测试不使用三级目次。

参考文献根据书后参考文献篇数的多少分为3个等级。A:20篇以内;B:20~40篇之间;C:40篇以上。

(3)将系统自动标引和自动分类得出的分类号,与手工标引给出的分类号比较。比较结果可分为4种情况(只比较主类号,不考虑复分号):①相同:分类号完全相符;②基本相同:分类号前三级相同但不完全相符;③不同:分类号完全不相符;④未分出:系统未能给出分类号。

对各标引源测评(打分)和分级,将结果填入测评数据库,例如第100号图书的平分和分级情况见表1。

表1 图书样例数据表片段

序号	书名	内容提要	一级目次	两级目次	参考文献	提要字数
100	1	1	3	2	A	B

3 数据的统计与分析

3.1 统计各标引源的人工打分

表2是对中文图书5个标引源人工打分测评情况的统计汇总,共有415条数据。

表2 图书各标引源测评打分分值统计

标引源	4	3	2	1	平均分值
书名	180	105	114	16	3.082
内容提要	166	93	133	23	2.969
一级目次	129	99	142	45	2.752
两级目次	145	96	137	37	2.841
参考文献	140	78	152	45	2.754

3.2 图书的内容提要和参考文献的统计

将图书的内容提要和参考文献分别作标引源进行自动标引和自动分类,再将分类结果与原分类号作比较可得到表3、4。

表3 内容提要自动分类结果统计

(其中X为内容提要字数)

分类结果	A级 $0<X\leq150$	B级 $150<X\leq250$	C级 $250<X\leq500$	D级 $X>500$
相同	29	100	44	12
基本相同	27	41	28	11
不同	33	61	43	17
未分	5	28	14	5
合计	95	230	129	45

表4　参考文献自动分类结果统计

(其中X为参考文献的条数)

分类结果	A级(0<X≤20)	B级(20<X≤40)	C级(X>40)
相同	91	49	47
基本相同	31	31	45
不同	49	40	64
未分	22	15	16
合计	193	135	172

3.3　统计数据的分析

根据表2,可以得出以上5个标引源主题表达能力的先后顺序,评判打分的分值越高,表明其主题表达能力越强。因此可以得到如下排列顺序:书名>内容提要>两级目次>参考文献>一级目次。

(1)表2表明,书名的主题表达能力是最强的。书名人工打分的平均分值为3.082,比其他各项标引源的平均分值都高,其他各项标引源平均分值均在3.0以下。当然也存在不能反映图书主题的情况,统计显示有16条数据不能反映主题内容,这是因为有些图书题不达义,书名与图书的主题不很相符。同时,统计数据表明,书名中能很好反映和基本反映图书主题的有285条,占68.67%。

(2)内容提要本来就是对图书主要内容的概括,应该能够比较正确反映图书的主题。统计也显示了内容提要的平均分值是2.969,仅次于书名;人工打分分值为4和3的分别有166条和93条,占总数的比例为62.41%,说明有60%以上的内容提要能很好反映图书主题或能基本反映图书主题。从各项标引源的平均分值来看,内容提要的平均分值比书名的低0.113,而比一级目次、两级目次、参考文献均高。因此,内容提要的主题表达能力是仅次于书名的。

(3)目次是图书的重要部分,总体概括了图书内容,也规定了图书的结构。从统计数据可以看出,一级目次和两级目次的平均分值分别为2.752和2.841,两级目次的主题表达能力强于一级目次。另外,两级目次的平均分值2.841低于内容提要的2.969,高于参考文献的2.754,两级目次的主题表达能力强于参考文献而低于内容提要。

(4)参考文献与一级目次的平均分值分别是2.754和2.752,两者相差仅0.002,主题表达能力持平;但一级目次中能很好反映主题的数目是129,明显低于其他各个标引源的数目。参考文献的主题表达能力强于一级目次。

(5)根据表3,将图书的内容提要按等级划分后,可以得出等级为B(字数为150~250)的是最多的,有230条数据。同时也对分类结果进行了统计,分类结果为"相同"且等级为B的最多,数目为100条。内容提要的字数在150~250之间主题表达能力较强,也比较有利于图书的正确归类。提要过短或者过长,都不利于自动标引和分类。

(6)根据表4,将参考文献按等级分类后,可以得到等级为A、B、C的数据条数分别是193、135、172,这说明参考文献的篇数在20以内的较多。再结合分类结果看,等级为A分类结果为"相同"或基本相同的是122篇占总数的71.3%,而等级为B、C的则分别为80、92,占总数的66.7%和59.0%。参考文献篇数在20以内的能较准确地反映主题,与主题相关性较大。等级为C的分类结果为"不同"的有64,是最多的,多于其他两个等级的49和40。这说明参考文献过多,在40篇以上的表达主题能力较弱。原因是参考文献过多,涉及的内容比较广泛,主题比较分散,主题词的抽取缺乏针对性。

4　加权方案设计及其结果分析

期刊、网页的加权标引设计已有专文讨论[8~9]。本文通过对图书的书名、内容提要、一级目次、两级目次、参考文献等主题表达能力的分析,确定了书名、内容提要、两级目次、参考文献作为图书自动标引的标引数据来源,并设计了如下4个标引加权方案:①书名+内容提要;②书名+内容提要+两级目次;③书名+内容提要+参考文献;④书名+内容提要+两级目次+参考文献。然后利用已经开发的自动标引与自动分类系统,按不同方案分别对数据进行处理,并对分类结果作出统计(见表5)和分析。

(1)表5表明,显然在4个方案中,方案①的分类准确率(即分准率,分类结果与人工分类相比为"相同"和"基本相同"的所占比例)高于其他方案,但是方案①的分得率是最低的,漏分率达18%。其他3个方案的分得率均高于方案①。

(2)从选取标引源个数多少看,当标引源的数量为两个时,其分准率为71.2%,而标引源个数增加为2和3时分准率反而下降。这些数据表明,书名和内容提要已经可以满足标引的需要,不必再费时费力地增加其他标引源。

表5 分类结果统计(单项测定)

分类结果	相同		基本相同		不同		未分		相同与基本相同的分类结果所占比例%
	数量	比例%	数量	比例%	数量	比例%	数量	比例%	
①书名+提要	196	39.20	96	19.20	118	23.60	90	18.00	71.2
②书名+提要+两级目次	194	38.80	104	20.80	134	26.80	68	13.60	69.6
③书名+提要+参考文献	200	40.00	100	20.00	147	29.40	53	10.60	67.1
④书名+提要+两级目次+参考文献	212	42.40	92	18.40	146	29.20	50	10.00	67.6

综合考虑分准率和分得率,采用情报检索评价时采用的检准率和检全率的调和值,即F1值测算(结果见表6)。

$$F1=\frac{2\times分得率\times分准率}{分得率+分准率}$$

调和测定值F1最高的方案④和最低的方案①,二者只相差1%。因此一般可以考虑采用标引源最少,检准率最高的方案①(只有书名+提要),既高效,又省力。数据处理时给各标引源(书名、内容提要、两级目次、参考文献)设置的权值可建议为5:3:2:2。

表6 分类结果统计(综合测定)

指 标	分准率%	分得率%	F1(综合值)%
①书名+提要	71.2	82	76.22
②书名+提要+两级目次	69.6	86.4	77.10
③书名+提要+参考文献	67.1	89.4	76.66
④书名+提要+两级目次+参考文献	67.6	90	77.21

5 结语

表6表明,与人工分类的结果相比,自动分类的分准率近70%,分得率已近87%,基于《中图法》知识库的自动标引和自动分类系统,用于图书是可行的。如果针对图书的特点对该系统稍加改进,自动标引和自动分类的效果将会更好。表6还表明,经过评测,选用数据源最少(书名+提要)的方案,与选用数据源最多的方案,其效果相差无几。因此,系统只选用图书的书名和内容提要(或简介)这两个最重要的标引源,即可完成自动分类。

限于时间,实验还存在一些问题:一是目前只选取了单一的经济类图书,未涉及其他学科,对测试结果会有一定影响;二是采用人工评测,人的主观因素有一定影响,今后应设法降低主观因素的影响;三是分类知识库还需完善,必须解决未登录词识别和及时补充的问题;四是测试学科范围和样本总数都较小,今后还需扩大学科范围,增加受试的文献数量。

参考文献

1 查贵庭,侯汉清.基于多词表的自动标引技术研究——新华社新闻稿自动标引的实验.情报学报,2002,21(3)
2 丁璇,侯汉清.中文网页标引源主题表达能力的调查统计.大学图书馆学报,2002(6)
3 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)
4 章成志,侯汉清.面向概念挖掘的文本层次模型研究.中国图书馆学报,2005,31(2)
5 薛鹏军,侯汉清.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(1)
6 赵妍等.中文期刊论文自动标引加权设计研究.新世纪图书馆,2004(1)
7,8 侯汉清,章成志,郑红.WEB概念挖掘中标引源加权方案初探.情报学报,2005,24(2)
9 薛春香,侯汉清.用于中文信息自动分类的中图法知识库的构建.中国图书馆学报,2005,31(5)

顾燕萍 江苏省昆山市高级中学图书馆助理馆员。通信地址:江苏昆山。邮编215300。

侯汉清 南京农业大学情报系教授。通信地址:南京市。邮编210014。

王晓红 天津农学院图书馆助理馆员。通信地址:天津。邮编300000。

(来稿时间:2006-03-13)