专著内容索引的编制方法探析
——以《情报检索语言与智能信息处理丛书》为例

于倩倩

【摘 要】文章以《情报检索语言与智能信息处理丛书》为例，介绍了专著内容索引的编制过程及方法。包括名称索引的编制过程、主题索引的编制过程及技术要点，并对《丛书》索引质量进行评析。

【关键词】专著 内容索引 索引编制 索引质量 索引评测

Abstract: This paper takes Information Retrieval Language and Intelligent Information Processing Series as an example to introduce the process and method of compiling monograph content indexes. Firstly, it introduces the process of compiling name index; then, it analyzes the process and technological points of compiling subject index; finally, it evaluates the quality of the Series indexes.

Key words: monograph content index index compiling index quality index evaluating

1 知识组织的重要方式——内容索引

在互联网迅速发展的今天，人们越来越重视基于内容的文本检索，知识组织、揭示方式具有举足轻重的地位。通过字符匹配的全文检索不能满足人们对知识的精准性需求，对文本内容进行主题标引是提高检索结果质量的重要途径。揭示文献内容、便于人们进行概念检索的主题标引分为两类，一类是浅标引，即只对文献内容进行整体标引，通常一本书最多标引 2 ~ 3 个主题词或词串，一篇期刊论文或学位论文最多标引 5 ~ 7 个主题词或关键词; 另一类是深标引，即对文献中的主要主题、次要主题及大大小小的主题，只要有检索意义，都可予以一一标引，标引深度达到几十、几百甚至几千上万。后者标引最细致、最费力、标引深度最大的，是图书、论文的内容索引(即主题索引) 。

内容索引是对书刊内容中所论及的主题概念、词语和事项进行全面揭示的一种检索工具，是对图书、期刊论文和学位论文等内容的深度揭示和组织。它将分散记载于大量文献中的知识单元组织起来，解决了目录、标引词(或称关键词) 只对文献作宏观著录和整体标引的不足，满足了用户对文献内容的微观揭示和检索需求，是知识组织的重要工具[1] 。在欧美国家，内容索引已经进入机读目录的特定字段，供人们进行“微内容”的检索。这种有别于全文检索的知识组织方式，越来越显示其独特的功能和价值，受到了广大用户的“青睐”。

专著索引是内容索引的一种，指以一部专著为对象的内容索引。为专著编制索引，一方面可以帮助读者较快地找到所需要的内容，另一方面读者通过浏览索引可以大体了解专著所论述的内容要点。本文总结《情报检索语言与智能信息处理丛书》(以下简称《丛书》) 索引的编制过程和经验，对专著索引的编制方法、技术等进行研究和探讨。《丛书》包括 8 本专著，著作的主题内容可以归结为“情报检索语言的自动化”和“自然语言检索”两个相关的问题，属于当前信息检索学术研究的前沿课题，对“信息检索自动化的升级问题”起了很好的开拓作用，为继续研究打下了基础[2] 。编著者为每部专著都配制了书后索引。《丛书》索引分为名称索引和主题索引两部分。名称索引收录人名、机构名、书名、系统名、数据库名、网络名等，主题索引收录文中论述的概念以及一些重要的文献名、系统名等。

2 名称索引的编制

名称索引的编制可划分为文本预处理、款目编制及排序、索引排版处理三个阶段。其中，文本预处理包括文本格式转换、机器分页和人工分页校订三个步骤，款目制作及排序包括手工勾标、机器抽词、款目拼接及排序等部分。名称索引的整体编制流程如图 1。

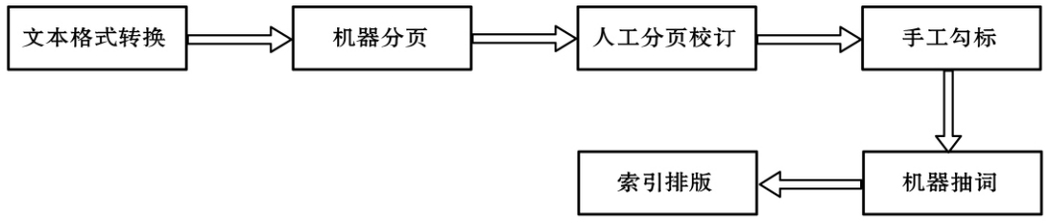

图1 名称索引整体编制流程

2.1 文本预处理

2.1.1 文本格式转换

首先将排版ps格式数据转换为pdf格式文本，然后通过Solid Converter PDF V4或其他相关软件将pdf格式文本转换为word格式版本。也可直接采用方正排版使用的小样文件（如．BAK文件或．FDB文件）。

在将pdf格式转换为word格式文本过程中，可能会出现字与字之间产生空格、英文字母转换成乱码等问题。对于字间空格现象，可利用word替换功能，将空格去除，以免影响词语匹配；对于英文字母转换成乱码问题的处理则需要对照pdf格式文本进行人工核查。

2.1.2 机器分页

《丛书》索引的标引范围决定不对各书的序言、目次、参考文献及后记等进行索引编制，故编制索引前先将文章中的参考文献删掉，以免造成自动抽词的错乱。将文后的参考文献删除后，缩小正文字体，以便在每页中留出足够空间添加机器标引的域标记，然后在正文中插入页码。

2.1.3 人工分页校订

机器添加页码完成后，要谨防页码出错，需要查看机器添加的页码与格式转换时生成的页码是否一致。若不一致，需要查找页码错乱原因，以确保索引编制的准确性。机器添加的页码与格式转换时生成的页码的一致性是十分重要的，它是保证索引质量的前提条件。页码错乱常见原因如有的书页上标引内容较多、版面空间不足等，这时，需要缩小正文字体或缩小行距，以便压缩篇幅，保证有足够的页面空间添加标记。

2.2 款目制作及排序

索引款目是组成索引的基本单元，一条索引款目记录专著中的一个索引对象并指出该索引对象在专著中的确切位置[3]。索引款目一般由标目和出处项组成，其中标目又分为主标目和副标目（或说明语）。主标目记录索引对象的基本内容，是索引款目排序也是读者检索的依据，副标目或说明语进一步限定主标目使之更具体化，出处项由页码组成。

名称索引款目制作宜于采用机器抽词法，即用word索引功能从文本中自动抽取，以手工勾标为辅。编制过程有两种方法，一种是先手工勾标出人名、机构名等系列名称，然后用word索引功能中的标记索引项逐项进行标记，最后生成名称索引，并对同义词予以规范；另外一种是标出人名、机构名等系列名称后，制作对照表（如表1），在对照表中对同义词进行规范，如果对照表中索引项较多，可将对照表分为几个子对照表，放置在不同的word文档中，然后利用word索引功能中的自动标记对各word文档中的索引项进行批处理，最后生成名称索引。至于文献名称，大多前后围以书引号，易于识别和机器抽取。有些文献没有添加书引号，则需要手工勾标，再用机器批处理。

表1 机器自动标引（批处理）用对照表

索引项	规范后的名称
国际十进分类法	国际十进分类法（UDC）
国际十进制分类法	国际十进分类法（UDC）
UDC	国际十进分类法（UDC）
都柏林核心集	都柏林核心元数据（DC）
Coates	Coates，E. J.
荷兰国家图书馆	荷兰国家图书馆

在对索引项进行标记过程中，无论使用哪种方法都应注意索引项的标记顺序，应该先长后短。如表 1 中的都柏林核心元数据（DC）应在国际十进分类法（UDC）之后进行标记，否则系统在对 DC 标记过程中会误标 UDC 中的 DC，导致标记错误，而再进行 UDC 标记时，由于被 DC 标记破坏，而导致 UDC 标记失败或遗漏。

规模较大的书籍所编索引往往设置参照项，由于本《丛书》每种专著索引篇幅不大，故不专门设置参照。对于同义词、译名及新的专有名词，处理如下：正式词在前，同义词在后，用括号括起；不同译名的标目，一般以中文为主，外文在后，用括号括起；有些新的系统名、网站名可以直接采用流行的英文缩写或缩略语，可以不附外文全名（往往太长）或中译名。对于外国人名，应当倒置，即姓在前，名或名的缩写字母在后，如 B. C. Vickery 应为 Vickery，B. C. ，B. Dialle 应为 Dialle，B. 。

2. 3 索引排版处理

索引排版包括索引的排版方式、所用字体、字号、分栏、缩进和间距等，都直接影响索引的美观大方和易用性[4]。如采用哪种排版方式？使用哪种字体、字号？分单栏、双栏或三栏？等。一般来说，索引排版要求清晰、易读、易查并且具有较高的密度。

在选择软件预定的排版格式如 word 插入索引中“来自模版”的排版格式后，计算机可自动生成一部完整的索引，然后导入 word 文档，设置字头（A/Z），添加眉线和眉题。常见的索引排版方式有分行式排版、连续式排版和表格式排版。《丛书》使用分行式排版，索引字体统一为小五号宋体，版面分两栏，每一栏里由两部分组成，左边为索引标目，右边为出处项。索引片段如下：


名称索引

名称索引

(本索引包括人名、机构名、网站名、文献名、系统名等)

A/Z（英文字母）
AGRIS，9
Aitchison，J.，23
CAB，9
CARMEN，108，123
CC，15
Chien，L.F.，128
CMIT，9
Coates，E.J.，5
CROSS，8
CWT，134，138-44
Dahlberg，I.，9
DDB，116，123

21
International Classification，9
ISKO，10
IZ，115-22
Krause，J.，108
LIBRIS，14
LIN，14
MACS，108
MARC，14，24，25，84，85
MEAD 数据中心，22
NEXIS 系统，22
Niehoff，R. T.，36
Northwestern 大学，18
OCLC，13，14

3 主题索引编制

主题索引采用手标机助法，即基本用手工逐页标引主标目和副标目，最后用机器补查，查看有无遗漏的论述。当然索引款目可以采用计算机排序。

3. 1 文本预处理

因为编制名称索引时已经对专著的文本进行了预处理和分页，因此编制主题索引时就不需要再对文本进行任何处理。

3. 2 编制过程与方法

3. 2. 1 人工勾标

人工勾标采用人工自由标引方式，即依据文献的内容与价值，选择能够表达专著某一局部内容的概念，可以大至专著的一个章节，也可以小至一个名词术语[5]，按照正文章节次序逐段进行勾标。选择的主题概念是否恰当，是决定索引质量的最重要环节。主题概念选取后，标目用词能否准确表达主题概念，也会影响索引质量的好坏。

一般来说，标目用词应当遵循表达被索引对象的语词尽可能准确专指、尽量采用专著原文中使用的词语和符合读者检索思路等规则[6]。

最后生成的索引不能等同于目次内容的字顺排列，不能简单地把大小章节标题全部做成索引款目。索引要以一种不同于目次页的方式来揭示专著的内容。正文中的图名、表名一般不宜直接拿来标引，图表名中的概念或专名可以适当进行标引。

3.2.2 机器处理

对于人工勾标出的标目，运用 word 里的查找功能，在转换生成的 word 底稿中进行查找，查看有无遗漏的地方，以确保索引编制的完备性。如果标目概念出现在图名、表名中，可以直接跳过，不做任何处理；有些页码可能只是提及标目概念，没有任何相关论述，对于这样的页码，不需要加入索引款目出处项。在主题索引编制过程中，只保留那些对某一主题概念进行论述的页码。

3.2.3 标目处理

(1) 单级标题式标目。只用一个词或词组作标目，不用副标目或说明语，如:

词素相似度计算，138－40，151－52

单向映射，74

对比体系，64，66，79

计算机文献标引对照系统，46－47

这种类型的标目比较简单但专指性相对较差。出处项一律标注所在页码，对于涉及多页的出处项，内容连续，则页码之间用连字符“－”表示，内容不连续，则页码之间用逗号分隔。例如，138－140，应写成 138－40；46－47 两位数不省略，此例不写成 46－7。

(2) 多级标题式标目（主标目＋副标目）。副标目对主标目涉及的范围按不同的研究方面加以限定。《丛书》索引中，“—”表示方面（顺读）。考虑到标目之间的层次关系对索引表达的影响，对于有些标目，在主标目之下除了副标目还设置了次副标目，一方面可以尽量挖掘各标目之间的联系，另一方面可以调节标目的专指度。主标目与副标目之间换行并空一字，注意副标目要前置一个连号（半个破折号）。如:

受控词表互操作，7－20，121－22

—必要性，9－10

—方法，18－20

—翻译法，20

—间接映射（动态映射），19

—派生法，20

—直接映射（静态映射），19

—自动匹配转换，18 —可视化，160－71

这种类型的标目用一个概括的词或词组作主标目，再用一个词或词组作副标目，限定或修饰主标目，可以使标目含义更加专指和明晰，族性检索性能好，但是副标目要求规范性，编制稍为复杂[7]。

(3) 倒置标题式标目。即主标目＋词组倒置部分，《丛书》索引中以 “;” 表示倒置（逆读）。如:

词汇转换语义关系，11－12

，部分等同，11

，广义等同，11

，完全等同，11

，完全不等同，12

，狭义等同，11

这种类型的标目是将词组标目开头的限定部分进行倒置，可以增强标目字面成族的效果，提高查全率。

(4) 带限义词的标目

限义词置于标目后用圆括号括起，主要用于区分标目中的多义词或同形异义词。有时标目太长也可以加限义词进行处理，它能够对标目做出必要而简略的补充和说明。它可以置于主标目之后，也可以置于副标目或倒置标题式标目之后。如:

关键词自动抽取方法（传统），126－28
—词库匹配法，127－28
—基于词频统计的关键词抽取法，126－27
—基于 N－gram 频率统计的方法，128
—完全 N－gram 标引法，127
—文法分析法，126

在《丛书》索引中，将同义词、简称或全称、原文或中译名等也置于标目后用圆括号括起，即将意义相同的标目统一合并为一个。一方面可以为用户提供更多的检索途径，另一方面可以省略参照。如:

XML（扩展标记语言），162－63
科图法（中国科学院图书馆图书分类法），124－25
一体化医学语言系统（UMLS），38－40
—超级叙词表，38－39
—词串标识符（SUI），39
—词语标识符（LUI），39
—概念标识符（CUI），39
—语义网络，39－40
—专家词典，39－40
中国图书馆分类法（中图法，CLC），22－25，84－86，148
—概况，148－49
—体系，150－52

其中，同义词的处理方式与名称索引中同义词的处理方式相同。正式词在前，同义词在后，用括号括起；不同译名的标目，一般以中文为主，外文在后，用括号括起。这种做法可以使主题索引同时担当译名表的作用。如:

最大似然估计法（极大似然估计法，LogL），91－94
—联合概率加权，93
—算法，92－94
—应用，93－94
—原理，91－923.3 索引排序

索引款目经过排序组织，才能产生检索功能。索引款目通过标目排序后形成索引，可以为读者提供某个具体索引款目在索引中的确切位置，从而使全部索引款目形成一个可检顺序，提高读者的查检速度[8]。《丛书》索引在提取主题概念的同时对页码进行了相应的归并排序，余之主要对主标目、副标目和次副标目进行排序，使用汉语拼音排序法，由计算机进行自动排列。由于目前没有单独的副标目排序功能，必须先还原（即主副标目连写）后排序，即先还原成如下形式:

检索语言的兼容模式，4－10
检索语言的兼容模式，标准化，4－5
检索语言的兼容模式，中介词典，5－6，26
检索语言的兼容模式，宏观词表与微观词表，6－7
检索语言的兼容模式，宏观词表与微观词表—基本思想，7
检索语言的兼容模式，宏观词表与微观词表—兼容方法，7
检索语言的兼容模式，集成词表，8－10

升序排列后，需人工进行整理，再删繁就简，结果如下:

检索语言的兼容模式，4－10
，标准化，4－5
，宏观词表与微观词表，6－7
—基本思想，7
—兼容方法，7

，集成词表，8 -10

，中介词典，5 -6，26

4 《丛书》索引质量分析

4.1 索引规模

索引的规模（详细、完善程度，即印刷篇幅）是索引设计中最主要的考虑因素之一。大多数图书索引的规模，是作者与出版社在经济上达成的“协商”方案。作者希望索引编得全面完善，能够充分满足读者需要，但出版社希望索引不要因为编得太细而导致成本过高[9]。

对索引规模的计算一般有两种方法。一种是“索引规模 = 索引的总页数/正文总页数”；另一种是“索引规模 =（索引总页数* 每个索引页的行数* 栏数）/（正文总页数* 每页正文的行数）”。我们使用第一种方法对8册《丛书》索引规模进行了统计，结果见表2。

表2 索引规模统计结果

正文页数		索引页数		索引规模
整部丛书	平均每册	整部丛书	平均每册	0.0684
1 243	155.375	85	10.625	

根据国外的经验，索引规模1%～3%属较短索引，3%～5%属中等规模索引，5%～7%及以上属编制精良的索引。从表1中可以看出，《丛书》的平均索引规模为6．84%，由于排版方面的稀疏，实际索引规模应比计算值要小，因此《丛书》的索引规模应属中等以上。

4.2 标引深度

标引深度又称标引网罗度或引得深度，是指对文献内容进行周详标引的程度，较高的标引深度可以为用户提供较多的检索入口，反映的主题内容较专指、较准确，检索的查准率可以得到提高。张琪玉[10]认为专著索引的标引深度以“索引篇幅/正文篇幅”计算，在本《丛书》中，标引深度具体到每页的索引款目数，即标引深度 = 索引款目总数/正文页数。由于有的索引款目指向章、节内容，有的标引地址含一条及以上，在本研究中统一记为一条索引款目。对整部丛书及每册书的标引深度的统计结果见表3。

表3 标引深度统计结果

序号	书名	作者	正文页数	索引款目数	标引深度
1	网络环境中知识组织系统构建与应用研究	薛春香	201	640	3.184
2	面向信息检索的汉语同义词自动识别	陆 勇	114	225	1.974
3	自然语言叙词表自动构建研究	杜慧平 仲云云	131	241	1.840
4	文本自动标引与自动分类研究	章成志 白振田	166	464	2.795
5	情报检索语言的兼容转换	张雪英	150	396	2.640
6	受控词表的互操作研究	戴剑波 刘华梅	170	384	2.259
7	领域本体的半自动构建及检索研究	何琳	190	239	1.258
8	基于引文分析可视化的知识图谱构建研究	李运景	121	263	2.174
	《丛书》		1243	2852	2.294

由表3可以看出，《丛书》标引深度分布于1～4个/页标引词，平均标引深度为2.294个，其中《网络环境中知识组织系统构建与应用研究》的标引深度最高。标引深度过小，会导致检准率下降，而标引深度过大，会增加索引的篇幅和编制成本，因此对标引深度要加以必要的控制。

4.3 索引密度

索引密度为索引正文每页所含的款目数量，即索引密度 = 索引款目总数/索引页数。同样将指向章节内容、含一条及以上地址的索引款目统一记为一条索引款目。索引密度往往取决于索引的款式和排版，直接影响用户每次浏览每个索引页获取的索引款目数，影响索引的使用效率。另外，在款目数相同的情况下，索引密度越高，则索引的篇幅越小，索引的编制成本也就越小[11]。对《丛书》索引密度进行统计，结果如表4。

表4 索引密度统计结果

序号	书名	索引页数	索引款目数	索引密度
1	网络环境中知识组织系统构建与应用研究	18	640	35.556
2	面向信息检索的汉语同义词自动识别	7	225	32.143
3	自然语言叙词表自动构建研究	7	241	34.429
4	文本自动标引与自动分类研究	12	464	38.667
5	情报检索语言的兼容转换	12	396	33.000
6	受控词表的互操作研究	11	384	34.909
7	领域本体的半自动构建及检索研究	9	239	26.556
8	基于引文分析可视化的知识图谱构建研究	9	263	29.222
	《丛书》	85	2852	33.553

从统计结果可以看出，《丛书》的索引密度在26～39之间，平均索引密度为33.553，其中《文本自动标引与自动分类研究》的索引密度最高。由此可以看出，即使预先规定排版格式，由于有的索引款目标目较长或具有多个出处项，使其占用两行或更多，导致每页索引中包含的款目数量减少，又或实际操作中的差异导致每本专著的索引密度不同。

5 结语

在这套丛书中，8本专著书后索引的款式和排版格式基本一致。这是因为在索引编制之前对数据处理方式、排版处理以及标目处理等做出了统一规定。由此可见索引设计对于索引编制的重要性。但是，这套丛书的索引仍存在一些问题。《丛书》中名称标目和主题标目应当尽量规范统一，然而由于8本专著对名词术语的使用不一，加之编制索引时对标目缺乏控制等原因，导致丛书中名称和主题标目的使用出现一定的差异，即标引不一致现象。另外由于校对的疏忽，致使丛书中主题索引的排版出现大面积的错误，因而以勘误的形式而重印。

长期以来，专著索引在西方出版界受到广泛重视，西方读者习惯于使用专著索引来检索自己所需要的内容。专著索引编制的好坏通常作为评判图书质量的重要指标之一。专著索引也是国际学术著作通行的惯例，是图书结构规范化和标准化的要求[12]。在我国，现代编辑出版很少编制专著索引，甚至在翻译引进西方著作时，为减少篇幅等原因将外文索引直接删除。近期国家新闻出版总署负责同志多次在重要会议上提出，“将在中国出版政府奖、国家出版基金和‘三个一百’原创出版工程的评审标准中增加一条：凡是索引、注释不规范的图书一律取消评审资格”[13][14][15]，这将为图书内容索引的发展带来新的契机。

注释

[1] 张琪玉．索引工作的性质与索引工作者劳动的性质．中国索引，2004（3）：2－3
[2] 侯汉清主编．情报检索语言与智能信息处理丛书．南京：东南大学出版社，2009：1－2
[3] [6] [7] [9] 张琪玉．图书内容索引编制法：写作和编辑参考手册．北京：化学工业出版社，2006：12－57
[4] [11] 李华．《中国图书馆分类法》第2、3、4版索引测评．图书馆建设，2005（1）：52－54
[5] [8] 侯汉清主编．索引编制手册．北京：中国标准出版社，2012：28－52
[10] 张琪玉．专著索引．江西图书馆学刊，2003（2）：1－2
[12] 黄远慧．浅议图书内容索引的推广．才智，2011（24）：89－90
[13] 北青网．第三届国家原创出版工程昨公布入围名单 新闻出版总署副署长邬书林称——图书索引不规范不得参评政府奖．北京青年报，2011－12－28：B11
[14] 庄建．对不规范的学术著作说“不”．光明日报，2012－01－12：9
[15] 张弘．总署将规范学术著作出版：没有索引不给资助，不参与评奖．新京报，2012－01－17：C11

于倩倩 中国科学院国家科学图书馆，中国科学院研究生院。