按照字符拼音排序系统的建立与编辑实践

李 杰[1)] 李瑞瑞[2)]

1)中国科技出版传媒股份有限公司,100009,北京;2)重庆科技学院法政与经贸学院,401331,重庆

摘 要 将字符按照拼音排序经常会出现在图书和期刊的编委姓名排序、参考文献排序、索引排序等场景中。由于汉字多音字的存在,现有软件无法实现根据场景自动确定多音字的准确读音并排序,导致排序结果必须进行人工审核。而且,现有软件的排序规则通常无法更改。笔者利用 VBA 代码实现多音字读音的手动选择,依托 Excel 工作表输入待排序条目、设置排序规则,以及输出排序结果,实现了按照拼音排序含有多音字的条目,且排序规则可以自定义。本系统适合在编委姓名排序、参考文献排序、索引排序等多种场景下使用,普及性、通用性较强,能够提高编辑工作效率。同时文章也指出了系统存在的不足之处。

关键词 字符拼音排序;多音字;VBA;排序系统

Establishment and editing practice of a new sorting system according to Chinese phonetic alphabet//LI Jie,LI Ruirui

Abstract Sorting characters according to Chinese phonetic alphabets often appears in scenarios such as sorting the names of editorial board members, reference list, and index of books and journals. Due to polyphonic Chinese characters, existing software cannot automatically determine the accurate pronunciation and sorting of polyphonic characters based on scenarios, resulting in manual review of sorting results. Moreover, the sorting rules of existing software often cannot be changed. The author uses VBA code to manually select the pronunciation of polyphonic characters, relies on Excel worksheets to input the items to be sorted, set sorting rules, and output sorting results, and achieves sorting of items containing polyphonic characters according to Chinese phonetic alphabet. The sorting rule can be customized. This system is suitable for use in various scenarios such as editor name sorting, reference sorting, and index sorting. It has strong popularity and universality and can improve editing efficiency. At the same time, the article also pointed out the shortcomings of the system.

Keywords sorting according to Chinese phonetic alphabets; polyphonic character; VBA; sorting system

First-author's address China Science Publishing & Media Ltd., 100009, Beijing, China

DOI:10.16811/j.cnki.1001-4314.2024.02.008

在日常编辑实践中,我们经常会遇到对一组内容中的字符按照一定规则进行排序的场景,如书刊编委姓名的排序、参考文献的排序,以及索引的排序等。目前,常用的排序规则包括按照笔画排序和按照拼音排序 2 种。在对书刊编委姓名进行排序时,常会使用按照编委姓氏笔画排序或姓氏拼音排序;对使用著者-出版年制稿件的参考文献进行排序时,GB/T 7714—2015《信息与文献 参考文献著录规则》要求先将参考文献按照文种集中,分为中文、日文、西文、俄文、其他文种 5 部分,然后按著者字顺和出版年排列[1];对图书索引条目进行排序时,常会按照条目字母顺序(即拼音)排列。有学者利用"贝特之姓氏笔画排序"软件实现人名按照姓氏笔画排序[2],笔者在实际工作中也经常使用,对书刊编委姓名排序帮助很大。也有学者借助 Excel 软件实现按照笔画顺序排列[3]或者拼音顺序排列[4]。但是,由于多音字的存在,按照拼音排序后,往往必须人工审核含有多音字的条目,以确保排序正确[4-7],这无疑为本就琐碎的编辑工作又添了一份负担。这种现象产生的原因,在于条目中的多音字在软件中默认按照一种固定读音参与排序。如果我们能够在软件排序前告知软件每个多音字的准确读音,然后再参与排序,那么排序结果自然就无需人工审核。为此,笔者在 Excel 软件中利用工作表存储字符读音与顺序,利用 VBA 代码建立一个事先手动指定多音字读音并且自动按照拼音排序的新型系统。该系统可以用于书刊编委姓名、索引、参考文献等排序场景,经实践检验具有方便快捷、排序规则可更改、出错率低的优点,无需对排序结果进行人工审核,从而提高了编辑工作效率。

1 整体架构

本系统的架构主要包括待排序条目录入与操作模块、排序结果显示模块和排序规则存储模块 3 个功能模块,见图 1。每个功能模块分别以 Excel 工作表或者 VBA 代码实现了待排序条目的录入、排序、显示,以及多音字标记和排序规则的存储与读取。

2 系统功能与特点

本系统的主要功能是,在使用者录入条目后,系统会从每个条目第一个字符开始,依次比较每个条目相同位置字符的拼音(对于汉字)或字母(对于英文单词)顺序,按照事先设置的字符顺序规则对条目进行重新排序。

本系统的特点如下:

1)使用者按条目依次输入需要按照拼音排序的

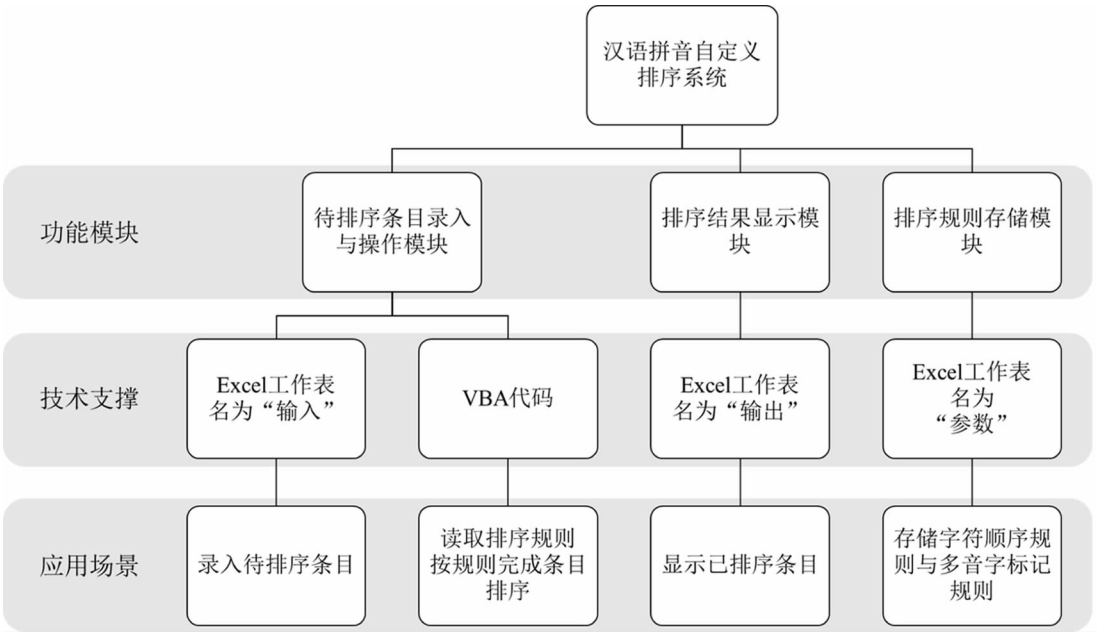


图 1　系统整体架构

内容后,系统会根据预先定义的字符顺序规则对条目的每个字符依次排序。预先定义的字符包括汉字、阿拉伯数字、希腊字母、罗马数字、西文字母(即大小写英文字母)5 类字符。对汉字,本系统收录了《通用规范汉字表》中的 8 105 个汉字[8],其顺序规则与 GB/T 13418—1992[9]一致:“按汉语拼音字母表顺序对汉字字符排列。如果拼音相同,比较音调,按阴平、阳平、上声、去声、轻声的次序对汉字字符排列。如果音和音调相同,比较汉字的总笔画数,从少到多对汉字字符排列。如果笔画数相同,比较汉字的起笔至末笔各笔笔形,依‘横、竖、撇、点、折’顺序排列。若起笔至末笔各笔笔形仍相同,则按汉字在国家标准汉字编码字符集中的编码值从小到大排列。”对阿拉伯数字,按照从 0 到 9 的顺序排列。对希腊字母,按照每个字母的大写与小写形式进行排序,即 A、α、B、β、Γ、γ、……。对罗马数字,按照每个数字的大写与小写形式进行排序,即Ⅰ、ⅰ、Ⅱ、ⅱ、Ⅲ、ⅲ、……。对西文字母,按照每个字母的大写与小写形式进行排序,即 A、a、B、b、C、c、……。

2)预先定义的 5 类字符之间的顺序可以自行更改,同时可以自行添加或删除新的字符或更改不同类字符的排列顺序。这样,当条目中包含不同文种时,其排序结果自然会按照不同类的字符排列。对参考文献,这样排序可以直接实现 GB/T 7714—2015 中“参考文献表采用著者 - 出版年制组织时,各篇文献首先按文种集中,可分为中文、日文、西文、俄文、其他文种 5 部分,然后按著者字顺和出版年排列”[1]的要求。对于索引,这样排序的结果也使得英文和中文开头的条目分别集中,只要调整各类字符的顺序就可以使排序结果符合 GB/T 22466—2008《索引编制规则(总则)》中汉字和非汉字字符混合出现时的排序:空格—序号—阿拉伯数字—罗马数字—拉丁字母(大写、小写)—日文假名(平假名、片假名)—希腊字母—俄文字母—汉字[10]。

3)条目中的多音字可以先手动选择读音,再自动参与排序。本系统对于多音字的控制方法操作简单,系统会根据规则条目判断待排序条目中是否含有多音字。对含有多音字的条目,系统会弹出交互窗体供使用者选择多音字的读音,然后根据使用者选择的读音参与排序。

4)使用前无需对排序条目中的特殊字符进行处理。在使用 Excel 对条目按照笔画排序时,有时需要事先对条目中的空格等字符进行预处理[3],以免这些字符影响排序结果。书刊编委姓名中经常出现的特殊字符是空格和圆点。索引条目中经常出现的特殊字符是各种标点符号,如短横线、圆括号、方括号、单双引号、书名号等。根据 GB/T 13418—1992 中的规定,缩写或简称中的圆点“·”、复合词中的短横线“-”、标点符号等属于非排序单元。本系统默认收录了一些不参与排序的特殊字符,见表 1。使用者可以对不参与排序的特殊字符进行自主管理。

5)本系统的规则建立在 Excel 工作表中,使得该软件的使用门槛很低,便于使用者自行增加、删除和修改规则。

6)本系统可对待排序的每个条目的前 10 个字符依次进行比较,当某个位置的字符相同时,会自动比较

表 1 系统默认不参与排序的字符

序号	字符名称	字符形式	Unicode 编码（十六进制）	备注
1	空格		0020	
2	中间点	·	00B7	
3	连字符	-	002D	
4	半角逗号	,	002C	
5	全角逗号	，	FF0C	
6	半角句点	.	002E	即小数点
7	全角句号	。	3002	
8	半角冒号	:	003A	
9	全角冒号	：	FF1A	
10	半角分号	;	003B	
11	全角分号	；	FF1B	
12	半角圆括号	( )	0028、0029	
13	全角圆括号	（ ）	FF08、FF09	
14	半角方括号	[ ]	005B、005D	
15	双引号	“ ”	201C、201D	半角与全角状态的前后双引号是同一个字符
16	单引号	‘ ’	2018、2019	半角与全角状态的前后单引号是同一个字符
17	书名号	《》	300A、300B	

下一个字符。在对编委姓名、索引、参考文献等内容进行排序的通常场景中，编委姓名字数通常最多为 3 字，也有少数民族人员姓名较长，超过 3 字的；索引条目有长有短，短为 1 字，长则 10 余字；参考文献条目则更长，前面几字多数为作者姓名或年代，如“张三，李四，王五，等. 2023. 论文题目[J]. 23(1):102 - 104”。因此，在排除逗号、句点、空格等字符后，通过前 10 个字符通常已经能够确定每条参考文献的先后顺序。如果遇到前 10 个字符都完全相同的条目，使用者可以修改该系统参与排序的字符数。

3 功能实现

3.1 待排序条目录入与排序操作

待排序条目的录入在“输入”工作表中进行，见图 2。

该表主要分为 A ~ C 3 个区域。A 区域由 2 列单元格组成，表头分别为“项目 1”与“项目 2”，用于输入待排序条目。为了避免“项目 1”字符过多影响运行效率，与排序无关却与“项目 1”连锁的内容可以放入“项目 2”列。如在图书编委姓名排序中，编委姓名是排序的内容，而编委单位与编委姓名连锁却不属于排序内容，由于编委单位的字符数通常较多，可以将编委姓名与编委单位分别放置在“项目 1”与“项目 2”列。B 区域为操作说明区，指导使用者正确操作。C 区域为排序按钮，点击即可对“项目 1”列的内容按照拼音排序。当“项目 1”中的字符包含多音字时，系统会依次弹出对话框，此时需要使用者手动选择多音字的读音，并点击“选择”按钮，即可完成多音字的读音选择，见图 3。

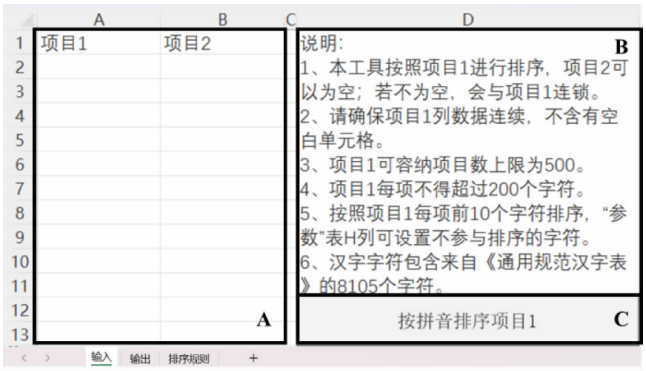


图 2 “输入”工作表的待排序条目录入与操作功能模块

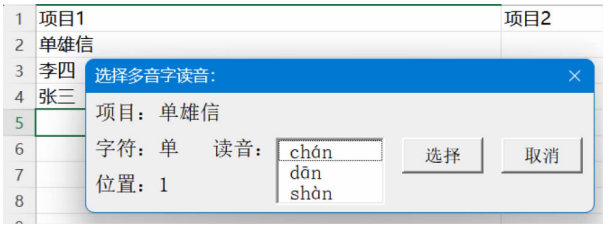


图 3 多音字读音的手动选择对话框

3.2 排序结果显示

排序后结果显示在“输出”工作表中，其结构与“输入”工作表完全相同，见图 4。只要点击“输入”工作表的“按拼音排序项目 1”的按钮后，系统会自动完成排序并跳转到“输出”工作表。

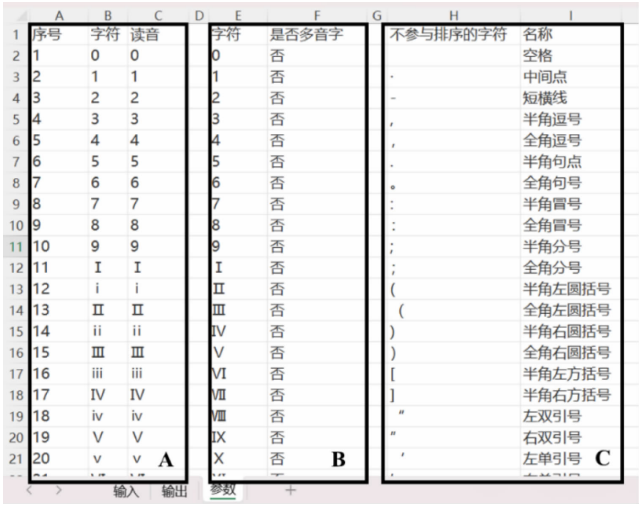
图 4 “输出”工作表的排序结果显示模块

3.3 排序规则管理

排序规则的管理在“参数”工作表中进行。该表包含 3 个区域——字符排序区(图 5 - A)、多音字管理区(图 5 - B)与不参与排序字符区(图 5 - C)。

3.3.1 字符排序区

字符排序区包括“序号”“字符”与“读音”3 列，这个区域规定了每个字符的序号与读音。其中，“序号”列为字符的排列顺序，用于排序时不同字符的比较，序号越小，排序越靠前；“读音”列为字符的读音，主要为区分多音字而设立供使用者选择，对于阿拉伯数字、罗



序号	字符	读音
1	0	0
2	1	1
3	2	2
4	3	3
5	4	4
6	5	5
7	6	6
8	7	7
9	8	8
10	9	9
11	I	I
12	i	i
13	Ⅱ	Ⅱ
14	ii	ii
15	Ⅲ	Ⅲ
16	iii	iii
17	Ⅳ	Ⅳ
18	iv	iv
19	Ⅴ	Ⅴ
20	v	v

字符	是否多音字
0	否
1	否
2	否
3	否
4	否
5	否
6	否
7	否
8	否
9	否
I	否
Ⅱ	否
Ⅲ	否
Ⅳ	否
Ⅴ	否
Ⅵ	否
Ⅶ	否
Ⅷ	否
Ⅸ	否
X	否

不参与排序的字符	名称
	空格
·	中间点
-	短横线
,	半角逗号
，	全角逗号
.	半角句点
。	全角句号
:	半角冒号
：	全角冒号
;	半角分号
；	全角分号
(	半角左圆括号
（	全角左圆括号
)	半角右圆括号
）	全角右圆括号
[	半角左方括号
]	半角右方括号
“	左双引号
”	右双引号
‘	左单引号

图 5　“排序规则”工作表的排序规则管理区

马数字、西文字母等不是多音字的字符,此列信息没有实际用处。对于汉字中的多音字,如阿(读音为 ā 或 ē,分别出现在 124 和 132 位置)会重复放置于此区中。不同读音的相同汉字的序号必然不同(见图 6),因此,只要获得某个多音字的正确读音,就能在此区中获得该多音字的唯一序号。当“项目 1”中每个待排序条目中的每个字符的序号都确定后,就能对不同条目间的第 1、第 2、第 3 个,直到第 10 个字符进行排序,最终确定不同条目的排列顺序。

121	120	y	y
122	121	Z	Z
123	122	z	z
124	123	吖	ā
125	124	阿	ā
126	125	啊	ā
127	126	锕	ā
128	127	腌	ā
129	128	啊	á
130	129	啊	ǎ
131	130	啊	à
132	131	啊	a
133	132	阿	ē
134	133	屙	ē
135	134	婀	ē
136	135	欸	ē

图 6　多音字“阿”在“字符排序区”的显示方式

3.3.2　多音字管理区

多音字管理区用于确定哪些字是多音字,包含“字符”与“是否多音字”2 列。这个区域的字符没有重复。本系统中收录了 597 个多音字。如果使用者需要增删多音字,可以在此区中操作,并在“字符排序区”相应的位置增删序号、字符与读音。

3.3.3　不参与排序字符区

不参与排序字符区用于存放不参与排序的字符,如空格、标点符号等,放入此区域的字符将在排序时被跳过,系统会使用下一个字符参与比较。使用者可以根据需要增加或删除此区域中的字符。

4　操作示例

4.1　含多音字条目排序示例

以阿胶、艾叶、巴豆、哈蟆油、蛤蚧、重楼这 6 味中药的拼音排序为例。它们的正确读音为阿胶(ē jiāo)、艾叶(ài yè)、巴豆(bā dòu)、哈蟆油(hā má yóu)、蛤蚧(gé jiè)、重楼(chóng lóu)。如果使用 Excel 排序,软件会默认将“阿”读音识别为“ā”、“蛤”读音识别为“há”、“重”读音识别为“zhòng”,导致排序结果错误。而使用本系统进行排序时,通过弹出窗体,使用者依次选择每个多音字的正确读音并点击“选择”按钮后,会得到正确的排序结果,见图 7。

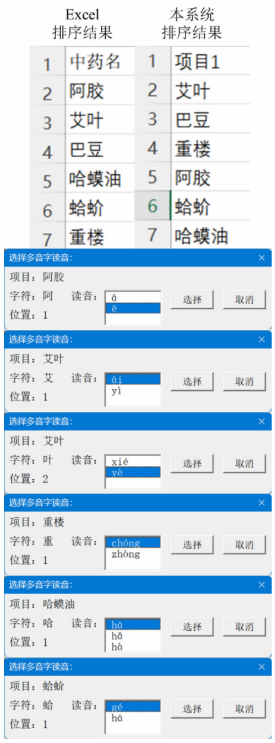


图 7　含多音字条目排序结果示例

4.2　含不参与排序字符条目排序示例

假设以下条目需要按拼音排序,见表 2。

如果按照 GB/T 22466—2008 中“汉字和非汉字字符混合出现时”[10]的排序要求来排列,我们需要在“不参与排序字符区”添加书名号“《”,使之不参与排序。然后录入表 2 中的条目,得到的排序结果见表 3。第 6 行“《药房和药品法》”条目从第一个汉字“药”开始参与排序,忽略了条目开始的“《”。

表2 待排序条目

项目1	项目2
药房法	Pharmacy
《药房和药品法》	Pharmacy and Medicines Act
药剂学	pharmaceutics
药品标识码	drug identification code
WHO 药品标准专家委员会	WHO Expert Committee on Specifications for Pharmaceutical Preparations (WHO-ECSPP)
精神药品	psychotropic substance
FDA 局长办公室	FDA Office of the Commissioner (OC)

表3 “《”不参与排序的排序结果

项目1	项目2
FDA 局长办公室	FDA Office of the Commissioner (OC)
WHO 药品标准专家委员会	WHO Expert Committee on Specifications for Pharmaceutical Preparations (WHO-ECSPP)
精神药品	psychotropic substance
药房法	Pharmacy
《药房和药品法》	Pharmacy and Medicines Act
药剂学	pharmaceutics
药品标识码	drug identification code

而当我们在“不参与排序字符区”删除“《”,那么“《”会参与排序,得到的排序结果见表3。第2行含有“《”的条目排在了最前面。

表4 “《”参与排序的排序结果

项目1	项目2
《药房和药品法》	Pharmacy and Medicines Act
FDA 局长办公室	FDA Office of the Commissioner(OC)
WHO 药品标准专家委员会	WHO Expert Committee on Specifications for Pharmaceutical Preparations (WHO-ECSPP)
精神药品	psychotropic substance
药房法	Pharmacy
药剂学	pharmaceutics
药品标识码	drug identification code

有时,我们还有更为特殊的要求,如汉字和非汉字字符混合的条目,排序时忽略非汉字字符,直接以汉字进行排列。那么,只需将西文字母添加在不参与排序字符区,然后再进行排列。这时,会得到如表5所示的排序结果:第3行条目“FDA 局长办公室”从“局”开始参与排序,排在以“精”开始参与排序的条目“精神药品”后面,忽略了条目开始部分的“FDA”;第8行条目“WHO 药品标准专家委员会”从“药”开始参与排序,排在以“药”开始参与排序的条目“药品标识码”后面,忽略了条目开始部分的“WHO”。

表5 “《”和西文不参与排序的排序结果

项目1	项目2
精神药品	psychotropic substance
FDA 局长办公室	FDA Office of the Commissioner(OC)
药房法	Pharmacy
《药房和药品法》	Pharmacy and Medicines Act
药剂学	pharmaceutics
药品标识码	drug identification code
WHO 药品标准专家委员会	WHO Expert Committee on Specifications for Pharmaceutical Preparations (WHO-ECSPP)

5 结束语

本系统以常用的 Excel 工作簿为载体,预先存储字符读音与顺序,对多音字的读音由使用者来选择,然后利用 VBA 代码对条目按照汉语拼音排序,可以实现含多音字条目的准确排序,并且可以自由设置排序规则,出错率低。得益于 Excel 软件的普及,使得该系统具有较强的通用性,能够在多种场景下迅速完成条目的排序,提高编辑工作的效率。当然该系统也存在一定的不足之处,比如每次排序可容纳的条目数量不超过500条,而且条目数量越多,系统的运行效率会有轻微的下降。有效缩短“项目1”列条目的字符数,尽量将不参与排序的信息放置在“项目2”列,可以最大限度地缩减系统的运行时间,提高运行效率。

6 参考文献

[1] 信息与文献 参考文献著录规则: GB/T 7714—2015[S]. 北京: 中国标准出版社, 2015
[2] 黎琳. 计算机在图书编辑加工中的应用和总结[J]. 科技传播, 2020, 12(9): 153
[3] 李学敏. 计算机类书稿图书编辑加工的步骤探究[J]. 新闻研究导刊, 2023, 14(11): 197
[4] 王超, 王乐. 引进类学术著作中英对照主题索引排序方法[J]. 科技与出版, 2013(8): 45
[5] 窦红娟. 谈百科全书辅文技术整理[J]. 出版参考, 2022(2): 81
[6] 宋慧敏. 著者-出版年制参考文献表编排过程中的排序问题[J]. 传播与版权, 2017(6): 60
[7] 于建辉, 刘新彦. 计算机技术在图书编辑加工中的应用: 以“工程机械系列双向词典”为例[J]. 传播与版权, 2021(12): 19
[8] 通用规范汉字表[M]. 北京: 中华人民共和国国务院, 2013
[9] 文字条目通用排序规则: GB/J 13418—1992[S]. 北京: 中国标准出版社, 1992
[10] 索引编制规则(总则): GB/T 22466—2008[S]. 北京: 国家图书馆出版社, 2012

(2023-11-25收稿;2024-01-14修回)