# 机读目录中文献版本关系识别与挖掘研究

## Research on Literature Editions Relationship Identification and Mining in Machine-Readable Catalogue

赵娅娜 （东南大学经济管理学院 江苏 南京 211189）
常 娥 （东南大学图书馆 江苏 南京 211189）

[摘 要] 实现同种文献不同版本数据的有效聚集，可以满足用户多重阅读与研究需求。在深入分析文献版本相关概念和常用文献版本资源聚集方法基础上，以中文机读书目数据为例，采用同一种文献不同版本数据归类与识别模型，并以 *Les trois mousquetaires* 为例，进行该文献不同版本资源的聚集与归类实验表明，所采用的文献版本资源识别模式能够较好地实现同一种文献不同版本资源的聚集，但对于题名变动较大或改换题名的同一种文献识别效果一般。

[关键词] 版本识别 书目数据 机读书目
[中图分类号] G250.1 [文献标识码] A

[Abstract] The effective clustering of different editions of the same literature could meet users' multiple reading and research needs. Based on analyzing related concepts of literature editions and clustering methods of literature edition resources, taking China machine-readable catalogue as an example, an experiment is carried out by a model of data classification and identification on different editions of the same literature. *Les trois mousquetaires* is selected as the experiment literature, and the result shows that the edition resources identification model could cluster different editions of the same literature, while it has little effect for those literature which change the title.

[Keywords] Edition identification; Bibliography data; Machine-readable catalogue

书目版本数据是对文献内容的重要揭示，有助于用户鉴别和选择不同版本的图书。近年来，由于出版行业的蓬勃发展，图书出版的数量和种类迅猛增长。2016 年，全国共出版图书 49.99 万种，较 2015 年增长 5.1%。其中新版图书 26.2 万种，增长 0.8%，重版、重印图书 23.8 万种，同比 2015 年增长 10.3%，占 2016 年出版总量的 48%[1]。这些图书进入图书馆后著录要涉及查重问题，同时需要识别同一种文献的不同版本，以便更好地进行文献的聚集。但由于版本信息复杂多样，版本的认定与著录本身比较困难，对于某一种图书版本的著录会出现不同形式的著录结果，究其原因可归纳为 3 种：第一，著录规则的不同，我国当前对于图书使用的著录规则主要有《中国文献编目规则》及其第 2 版、CALIS《中文图书著录规则》和《GB/T 3792.2 85/2006 普通图书著录规则》，不同的规则在对版本进行著录时有不同的要求 [2]。第二，规则的变更以及规则描述的不清晰造成编目员理解的不一致，如对于说明著作内容特征的文字（缩写本、绘画本、英汉对照本等），《中国文献编目规则》第一版中规定著录于版本项，而《中国文献编目规则》（第 2 版）规定一般不著录于版本项 [3]。第三，对反映图书版本的著录信息源前后描述不一致，如“中国分类主题词表（第二版）/ 国家图书馆《中国图书馆分类法》编辑委员会编”，版权页为“2005 年 9 月第 1 版”的著录 [4]。版本信息著录结果的不同，使得同一种图书分散在各个地方，难以将各个不同版本的图书进行聚集，读者在利用书目系统查询图书时，只能检索出同一题名的图书。对于该书的不同版本资源，需要读者自己识别，因而降低了读者对书目查询系统的使用体验。而对于图书编目人员而言，在对新入藏的图书进行编目时，影响查重效率，造成同书多编，也不利于从历史书目版本信息的著录中发现对版本著录存在的问题及进行更好的优化。

本文拟在深入分析文献版本相关概念和常用文献版本资源聚集方法基础上，以中文机读书目数据为例，提出同一种文献不同版本数据归类与识别模型，从而实现在已有机读书目数据中同一种文献不同版本资源的识别与聚集，以满足普通用户和专业人士的多重阅读与研究需求，具有重要现实意义。

ISSN 1004-325X | CN23-1331/G2

## 1 版本相关概念阐释

文献版本研究由来已久。西汉伊始，人们就讲究传本，把手抄的书籍称为“本”，每一代传人整理的书籍称为“传本”。进入宋、元，雕版印刷术盛行，把印刷的书籍称为“版”，版本由此而来。版本研究目的在于考究文献版本优劣，发现善本。随着现代社会出版技术的发展与进步，文献版本越来越丰富，需要对不同版本书籍进行更好的编目与聚集，以满足人们多重阅读与研究需求。要实现同种文献不同版本的聚集，需要了解什么是同一种文献以及什么是同一种文献的不同版本。

对于同一种文献，王玮琦认为，“同一责任者所写的同一题材的书为同种书。允许书名改变，但题材内容和第一责任者不能变。”[5] 马兰芳提出，“当题名、主要责任者、主要内容相同时被视为同一种书。”[6] 郁德祥认为，“同类中凡书名相同、内容关联、书名关联、内容相同者被认为是同种书。”[7] 这些学者对于什么是同一种书，都认同内容必须相同，而对于题名、责任者、出版、印刷等项是否相同则有不同的见解。本文认为由于第一责任者对文献的创作负有主要责任，若第一责任者不同，说明文献内容有可能发生了较大改动，可作为不同种文献处理。所以同一种文献是指同一责任者所著的主要内容相同的图书，包括不同版次、不同译者、修改者等改变了书名和体裁，但没有改变第一责任者和主要内容的文献。

对于同一种文献的不同版本的认定，马兰芳认为，题名、责任者、内容完全相同时为同一版本，主要责任者和体裁不变，内容有所变动时作不同版本处理[6]。《中国文献编目规则》(第二版)指出，版本是同一种文献(出版物)，因编辑、排版、装订或制作形式的不同而产生的不同的本子[8]。《普通图书著录规则图例手册》指出，出版本是指内容或形式上与原版有所不同的作品[9]。从对不同版本的认定可以看出，对于内容改变必然会产生不同的版本，但是对于形式改变是否应该认为是不同的版本，说法不一。因随着图书出版数量增加，其出版形式复杂多样，在一定程度上出版形式的改变能影响用户对图书的选择。所以本文认为当同一种文献的内容或形式发生改变时，都应该认为是同一种文献的不同版本。

## 2 常用文献版本资源聚集方法

图书馆在著录书目时，意识到了多版本问题的存在，为了将文献进行聚集，不同的学者提出了不同见解。赵伯兴提出，通过采用规范控制、分类和字段连接方法来聚集翻译作品[10]；王玉梅提出，字段连接法、统一标目法和归类一致法能够实现对于原作与译作、译作与译作的聚集[11]；但梁美宏认为，赵伯兴和王玉梅提出的版本聚集方法只能揭示两两之间的关系，关联程度较低[12]。所以在此基础上提出采用关联分析的方法来实现不同文献版本的深度关联。目前常用的版本资源聚集方法共有4种：款目连接法、规范控制法、同一分类法和关联分析法。

款目连接法主要指利用机读目录中相应字段实施控制。在中文机读目录中，主要通过451、452、453、454、455、456字段实现同一载体版本、不同载体版本，原作和译作以及原作和复制品的书目连接。4--字段作为选择性字段，虽然能实现书目连接，但实际应用操作难度较大，基本不做或很少做著录。另外我国中文文献用CNMARC著录，外文用MARC21，实现中文和外文之间的连接较为困难，所以难以实现多版本文献聚集。

规范控制法是指对图书题名和作者名称进行规范控制来实现书目连接。中文机读目录通过500字段的统一题名进行题名规范，作者名称控制则建立人名规范挡。但因目前绝大部分图书馆自动化系统都没有提供或者启用规范及挂接功能[13]，所以利用规范检索点发现同一种文献的不同版本的优势实际上无法体现。另外虽然采用500字段对题名进行规范，但在实际操作中对于统一题名的选择有较高要求。

同一分类法是指通过赋以统一的分类号和种次号，并且附加辅助区分号，在一定程度上能够实现不同版本资源聚集和区分功能。但是著录规则不统一及著录人员对版本信息源理解的不一致等原因，导致这一方法仍未发挥出该有的作用。而且在实际书目检索系统中，用某一分类号进行检索，往往得到的是这一类书，鲜有包含同一种文献的所有不同版本资源。

关联分析法主要借用关联数据名称唯一性特点，构建了基于关联数据和书目数据的文献版本关系发现方法[12]。目前该方法只研究了单属性版本关系发现，即只有一种属性不同，其他版本属性都相同的同种文献版本关系，对于多属性版本关系识别研究还未开展，如同版次不同版式关系、同版式不同版次版本关系、不同版次不同版式版本关系等。此外，通过关联分析法实现图书馆中不同版本书目

数据的聚集，首先要求将 MARC 中的书目发布为关联数据，而当前图书馆发布关联数据集较多的国家为德国、美国、英国、法国以及一些国际联合项目，亚洲只有日本发布了关联数据 [14]。

综上所述，这 4 种方式在一定程度上能够实现同一种文献不同版本资源的聚集，但在实际应用中作用不明显。为了便于读者在图书的不同版本资源中进行选择利用，本文尝试在已有的机读书目数据中进行多属性版本文献的自动识别与挖掘。

## 3 文献版本类型及在机读数据中的表现

### 3.1 文献版本类型归类

王玉梅提出文献版本有多种表现形式，载体不同时会产生不同的版本，载体相同时对于普通图书来说主要有原本与译本、不同名称、不同出版者、不同译者、收录丛书不同、影印本、注释本 7 种 [11]；房亚玉根据实际工作中遇到的情况将版本类型归纳为版次、印次、出版单位、装帧（装订、版式、册次）、不同译者、不同题名 6 种；何云、黄久斌通过对《普通图书著录规则图例手册》的理解，将版本归纳为版刻、版次、文种、文体、出版者等 13 个类型 [15]；梁美宏、曾建勋结合当前研究和实际编目情况归纳出版次、其他责任者、版刻、版式、装帧形式、语言、卷册、出版社、出版时间和书名共 10 种类型 [13]。

对于版本的类型各个学者看法不一，本文结合现有研究同时根据对《普通图书著录规则图例手册》的理解，从形式和内容上归纳出 12 种版本类型的划分依据，详见表 1 所示。

表 1 版本类型划分依据及内涵

| 种类 | 版本类型划分依据 | 含 义 |
|---|---|---|
| 形式改变 | 版刻 | 图书制版类型和复印方法，如手抄本、木刻本、油印、铅印、数字版本 |
| | 版式 | 图书排版方式，横排改为竖排，字体、字号和插图的变化等 |
| | 出版单位 | 图书出版单位的不同 |
| | 卷册次 | 原版书，有时会分成多卷或多册出版 |
| | 收录丛书 | 同一本图书在出版时，收录进不同的丛书类型 |
| | 装订 | 根据不同需求进行装订，主要分平装、精装、软精装、豪华装等 |
| 内容改变 | 版次 | 图书排版的先后次序，以标明图书内容经过修改、增补等较大变化。如第二版、修订版等 |
| | 其他责任者 | 第一责任者相同时，因译者、注者、点校者、改写者不同 |
| | 题名 | 同一作品不同翻译题名及以不同题名形式出版的同一作品 |
| | 文种 | 中文书名但正文是其他文种，如题名是中文，但正文是少数民族文字 |
| | 文体 | 从文言文到白话文的转变就属于文体的变化 |
| | 说明著作内容特征 | 对于图书的内容具有说明性的意义，如缩写本、节本、通俗本等 |

### 3.2 机读数据中版本类型对应的相关字段——以 CNMARC 为例

因文献版本类型复杂，以上所列的版本类型在机读数据中并不能一一对应，本文就机读目录中与版本相关的字段进行了整理，其对照关系如表 2 所示。

表 2 中文文献版本关系识别字段

| 属性 | CNMARC 字段 | 对应版本类型 |
|---|---|---|
| 国际标准书号 | 010 $b 装帧 | 装帧 |
| | 010 $d 册次 | 册次 |
| 文献语种 | 101$a 正文、声道语种 | 语种 |
| | 101$c 原著语种 | 语种 |
| 题名与责任说明 | 200 $g 其他责任说明 | 其他责任者 |
| 版本说明 | 205 $a 版本说明 | 版次 |
| 出版发行项 | 210 $a 出版发行地 | 出版单位 |
| | 210 $c 出版发行者名称 | |
| | 210 $d 出版发行时间 | |
| 丛编项 | 225 $a 丛编题名 | 丛编 |

## 4 文献版本关系识别与挖掘模型——以 CNMARC 为例

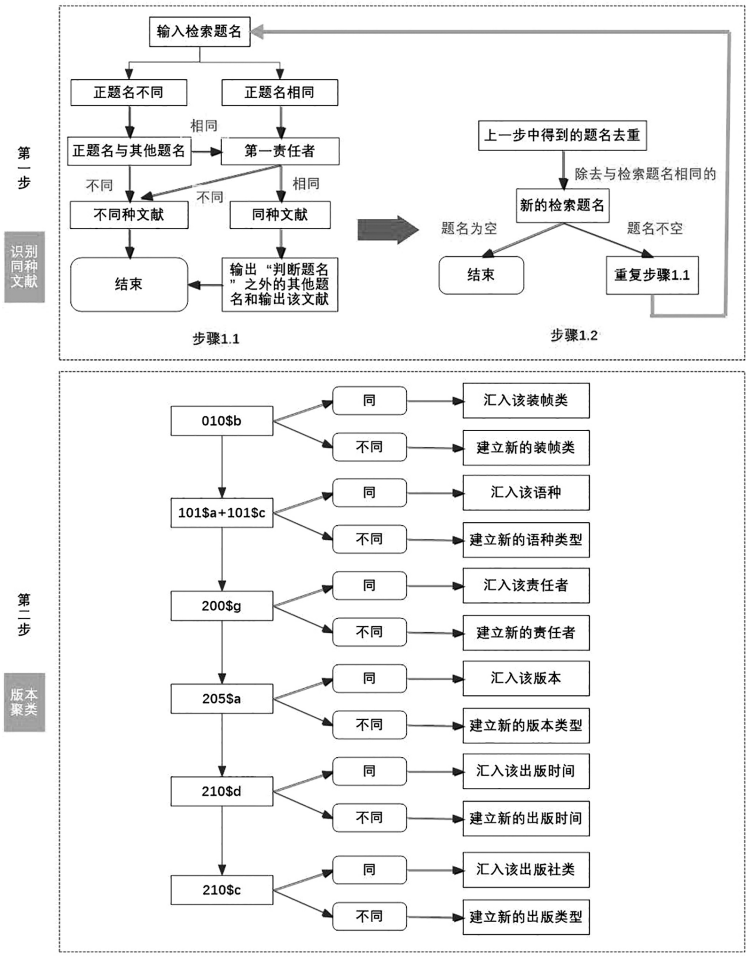
综上所述，本研究认为同一种文献是指同一责任者所著的主要内容相同的文献，包括由于不同版次、不同译者、不同修改者等改变了书名和体裁，但没有改变第一责任者的文献。当同一种文献的内容或形式发生改变时认为是同一种文献的不同版本。本文以 CNMARC 数据为例，构建了文献版本关系识别与挖掘模型，如图 1 所示。

由图 1 可知，文献版本关系识别主要包含两个步骤。

第一步，识别出某一种文献的所有书目数据，主要采用“题名 + 第一责任者”进行挖掘。

ISSN 1004-325X | CN23-1331/G2

图 1　文献版本关系与挖掘模型




说明：CNMARC 在著录书目数据时，与题名相关的信息著录在 200$a 正题名子字段和 5-- 相关题名块字段，责任者字段著录在 200$f 第一责任者子字段、200$g 其他责任者子字段和 7-- 知识责任者块。因此选用 5-- 字段的内容和 200$a 作为识别文献题名 200$f 的字段。200 子字段中的第一责任说明和其他责任说明在 7-- 字段会重复著录，为了做一简化处理，选用 200 子字段的信息。经过处理后 CNMARC 中中文图书著录项目中识别同一种文献相关的字段及子字段（见表 3）。

具体处理流程包含步骤 1.1 和步骤 1.2 两步。

步骤 1.1，给出原文献的题名，若检索出的文献正题名 200$a 与原文献题名相同时，继续比较与原文献 200 $f 第一责任说明是否相同，若相同，则结束匹配，认为这是同一种文献，然后输出“判断题名”之外的其他题名和输出该文献；若检索出的文献正题名 200$a 与原文献不同，则取正题名之外的其他的题名与原文献的题名比较，若相同，则继续比较与原文献 200 $f 第一责任说明是否相同，若相同，则结束匹配，认为这是同一种文献，然后输出“判断题名”之外的其他题名和输出该文献，若与原文献 200 $f 第一责任说明不同，则结束匹配，认为这不是同一种文献。

步骤 1.2，从步骤 1.1 中得到的所有题名进行去重，除去与检索题名相同的题名，形成新的检索题名集，若新的题名集为空集，则结束检索同种文献，若非空，则进入步骤 1.1 检索同种文献。

第二步，将步骤一中输出的所有文献进行去重后进入第二步的版本归类，将识别出的书目数据按属性值匹配的方式归入不同的版本类型。

说明：根据 CNMARC 书目版本数据中发现与版本相关的字段类型和实际聚类需求，将版本类型分为装帧、语种、其他责任者、版次、出版发行者、出版发行时间 6 种类型，按 CNMARC 与版本相关的属性逐项匹配，归入所属类别，因每个版本数据有多个属性，可按版本的性质归入多个类别。

首先将第一步识别出的所有文献的

表 3　中文文献同种文献识别字段

| 属性 | CNMARC 字段 |
|---|---|
| 题名与责任说明 | 200 $a 正题名 |
| | 200 $f 第一责任说明 |
| 相关题名块 | 500 $a 统一题名 |
| | 510 $a 并列题名 |
| | 512 $a 封面题名 |
| | 513 $a 附加题名页题名 |
| | 514 $a 卷端题名 |
| | 515 $a 逐页题名 |
| | 516 $a 书脊题名 |
| | 517 $a 其他题名 |
| | 518 $a 现代标准书写题名 |
| | 540 $a 编目员补充的附加题名 |
| | 541 $a 编目员补充的翻译题名 |

010$b 的值与原版文献比较，若取值相同，则归入该装帧类，否则建立新的装帧类别。不同版本语种的识别是通过 101$a + 101$c 来进行匹配的。其他责任者识别是将 200$g 与原版书目信息匹配，若相同，则归入这一责任者数据集，否则建立新的数据集，实现不同的注者、译者、校者的书目数据的聚类。版本识别是通过 205$a，匹配，得到同一种文献的修订版、增订版、第二版等。出版社的识别主要是 210 $c 数据与原版文献的匹配，聚类得到不同出版单位出版的同一种文献。时间识别是 210$d 出版发行时间来聚类，可以得到按不同时间出版的同一著作的不同版本。

## 5 文献版本关系识别与挖掘实证

### 5.1 数据来源

以东南大学图书馆书目数据为数据源，针对法国作家 Dumas Alexandre 的作品 *Les trois mousquetaires* 的不同版本文献资源进行识别与挖掘实验。其原著为法语版，有原版和续版。原版法语名称有“Les trois mousquetaires”“Trois mousquetaires”，英文译名有“Three musketeers”，中文译名有“三个火枪手”“三剑客”“三剑侠”“侠隐记”等。续版法语名称为“Vingt ans apres”，中文译名有“二十年后”“三个火枪手续集”“三剑客续”等。

### 5.2 文献版本关系识别过程

第一步：同种文献的判断。

通过对法国作家 Dumas Alexandre 的作品 *Les trois mousquetaires* 进行版本挖掘，共经过三轮检索与判断：第一轮，以题名“三个火枪手”检索，检出 24 篇文献，判别出同种文献 23 篇，第二轮，以“二十年后”“效忠国王”等 7 个题名进行检索，检出文献 34 篇，判别出同种文献 28 篇，第三轮，以“三剑客”“Vingt ans apres”等 4 个题名进行检索，检出文献 20 篇，判别出同种文献 16 篇，每轮的检索题名根据识别模型获得。将三轮获得的文献根据 001 字段的值是否相同（001 字段值具有唯一性），进行去重后获得该作品的同一种文献有 38 种。由于 MARC 数据源本身存在重复编目、错编等问题，所以对 38 种文献数据再进行人工清洗后，最终剩余 33 种文献数据。详细过程如下：

第一轮判断：本文以 *Les trois mousquetaires* 最常用的中文译名之一“三个火枪手”为题名进行检索，共检出 24 篇中文数据文献，分别编号为 A1–A24，提取出每条数据的正题名、其他题名信息和第一责任者，分别判断每篇文献与原著是否为同一种文献。

以 A1 为例，判断其是否为同一种文献：

（1）判断题名是否相同　A1 题名：三个火枪手 = 检索题名：三个火枪手，正题名相同

（2）判断作者是否相同　A1：第一责任者 =（法）大仲马，作者是 Dumas Alexandre 的中文译名

（3）判断是否是同种文献 是

（4）因为判断题名为“三个火枪手”，输出判断题名之外的其他题名：Trois mousquetaires，同时输出 A1。同理判断文献 A2–A24，结果如表 4 所示。

**表 4　第一轮检索文献判断情况及去重后的题名**

| 序号 | 同种文献 | 析出题名 | 序号 | 同种文献 | 析出题名 | A1–A24“析出题名”去重 |
|---|---|---|---|---|---|---|
| A1 | 是 | Trois mousquetaires | A13 | 是 | 无 | 1.Trois mousquetaires<br>2.Les trois mousquetaires<br>3.Three musketeers<br>4.The three musketeers<br>5. 二十年后<br>6. 效忠国王<br>7. 火枪手 效忠 国王 |
| A2 | 是 | 无 | A14 | 是 | 二十年后 | |
| A3 | 是 | Trois mousquetaires | A15 | 是 | 无 | |
| A4 | 是 | Les trois mousquetaires | A16 | 是 | 二十年后 | |
| A5 | 是 | Les trois mouspuetaires | A17 | 是 | The three musketeers | |
| A6 | 是 | 无 | A18 | 是 | 无 | |
| A7 | 是 | Trois mousquetaires | A19 | 是 | 无 | |
| A8 | 是 | Trois mousquetaires | A20 | 是 | 无 | |
| A9 | 是 | Les trois mousquetaires | A21 | 是 | 无 | |
| A10 | 是 | 无 | A22 | 否 | Three musketeers | |
| A11 | 是 | 无 | A23 | 是 | Les trois mousquetaires、效忠国王、火枪手 效忠 国王 | |
| A12 | 是 | 无 | A24 | 是 | 无 | |

注：表中的析出题名即判断题名之外的其他题名。

第二轮判断：以第一轮判断去重后的 7 个题名为检

ISSN 1004–325X | CN23–1331/G2

索题名进行第二轮检索，共检出 34 篇中文数据文献，将检出文献编号为 B1–B34，提取出每条数据的正题名、其他题名信息和第一责任者。与第一轮检索结果判断原理相同，依次判断文献 B1–B34 是否为同一种文献，结果如表 5 所示。

第三轮判断：以第二轮判断去重后的 4 个题名为检索题名进行第三轮检索，共检出 20 篇中文数据文献，将识别出的文献编号为 C1–C20，提取出每条数据的正题名、其他提名信息和第一责任者，同理判断 C1–C20 是否为同一种文献，结果如表 6 所示。

因第三轮去重后题名为空集，所以结束检索。

结论：共检出 78 篇，进行判断后发现同一种文献 67

**表 5　第二轮检索文献判断情况及去重后的题名**

| 题　名 | 序号 | 同种文献 | 析出题名 | 题　名 | 序号 | 同种文献 | 析出题名 | B1-B34“析出题名”去重 |
|---|---|---|---|---|---|---|---|---|
| Trois mousquetaires | B1 | 是 | 三个火枪手 | （the）Three musketeers | B18 | 是 | 三个火枪手 | 1. 三剑客<br>2.Vingt ans apres<br>3.《三个火枪手》续集<br>4. 三剑客续 |
| | B2 | 是 | 三剑客 | | B19 | 是 | 三剑客 | |
| | B3 | 是 | 三个火枪手 | | B20 | 是 | 三个火枪手 | |
| | B4 | 是 | 三个火枪手 | 二十年后 | B21 | 是 | 无 | |
| | B5 | 是 | 三个火枪手 | | B22 | 是 | Vingt ans apres | |
| | B6 | 是 | 三个火枪手 | | B23 | 否 | 无 | |
| Les trois mousquetaires | B7 | 是 | 三剑客 | | B24 | 否 | 无 | |
| | B8 | 是 | 三个火枪手 | | B25 | 否 | 无 | |
| | B9 | 是 | 三剑客 | | B26 | 否 | 无 | |
| | B10 | 是 | 三剑客 | | B27 | 是 | 无 | |
| | B11 | 是 | 三剑客 | | B28 | 是 | 《三个火枪手》续集 | |
| | B12 | 是 | 三个火枪手 | | B29 | 是 | 无 | |
| | B13 | 否 | 无 | | B30 | 是 | 《三个火枪手》续集 | |
| | B14 | 否 | 无 | | B31 | 是 | Vingt ans apres、三剑客续 | |
| | B15 | 是 | 三个火枪手、效忠国王、火枪手效忠国王 | 效忠国王 | B32 | 是 | 无 | |
| | | | | | B33 | 是 | Les trois mousquetaires 、火枪手效忠国王、三个火枪手 | |
| （the）Three musketeers | B16 | 是 | 无 | 火枪手效忠国王 | B34 | 是 | 三个火枪手、Les trois mousquetaires、效忠国王 | |
| | B17 | 是 | 三个火枪手 | | | | | |

**注：**表中的析出题名即判断题名之外的其他题名。

**表 6　第三轮文献判断情况及去重后的题名**

| 检索题名 | 序号 | 同种文献 | 析出题名 | 检索题名 | 序号 | 同种文献 | 析出题名 | C1-C20“析出题名”去重 |
|---|---|---|---|---|---|---|---|---|
| 三剑客 | C1 | 是 | Trois mousquetaires | 三剑客 | C11 | 是 | 无 | 无 |
| | C2 | 是 | Les trois mousquetaires | | C12 | 是 | Three musketeers | |
| | C3 | 是 | Trois mousquetaires | | C13 | 是 | 无 | |
| | C4 | 是 | Les trois mousquetaires | | C14 | 否 | 无 | |
| | C5 | 是 | Les trois mousquetaires | Vingt ans apres | C15 | 是 | 二十年后 | |
| | C6 | 否 | 无 | | C16 | 是 | 二十年后 | |
| | C7 | 是 | 无 | 《三个火枪手》续集 | C17 | 是 | 二十年后 | |
| | C8 | 否 | 无 | | C18 | 是 | 二十年后 | |
| | C9 | 是 | 二十年后 | 三剑客续 | C19 | 是 | Vingt ans apres 、二十年后 | |
| | C10 | 是 | 无 | | C20 | 否 | 热恋中的达达尼昂 | |

**注：**表中的析出题名即判断题名之外的其他题名。

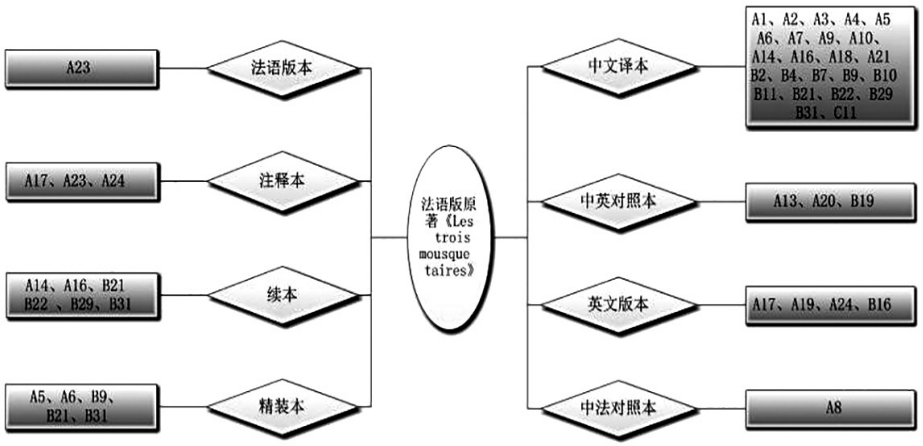
篇。因文献的 CNMARC 的 001 字段具有唯一性，所以根据 001 字段去重，如前所述，由于 MARC 数据源本身存在重复编目、错编等问题，所以经过人工清洗后最终获得与原著构成同种文献关系的 33 种文献（详细列表省略）。

第二步：对 33 种文献根据版本类型进行归类。

提取 33 种文献的相关 CNMARC 字段，进行版本类型归类，包括 200$a、010$b、010$d、101$a+101$c、200$g、205$a、210$d 和 210$c 这 8 个字段信息，部分信息如表 7 所示。200$a 表示正题名；010$b 表示文献的装帧类型，包括精装、平装等；010$d 表示文献的分册出版情况；101$a 表示文献正文所用语种，101$c 表示原著语种；200$g 表示其他责任者，意为除第一责任者外，对文献负有其他责任的人，包括译者、改写者等；205$a 表示版本类型，如第 2 版、修订版等；210$d 表示出版时间；210$c 表示出版单位。

本文根据表 1 给出的 12 种版本类型划分依据，并结合 33 种同种文献的版本信息，进行归纳总结后，将其划分为 8 种版本类型，包括中文译本、中英对照本、中法对照本等，具体版本类型及对应文献信息如图 2 所示。

图 2　文献版本类型关系




中文版本：中文版本的图书涉及 24 种，汇集了包括“三个火枪手”“三剑客”“二十年后”等为文献正题名的图书，其中一部著作分两册出版的图书有 14 种，以一册形式出版的图书有 10 种。图书出版年从 1978—2017 年，其出版社种类繁多。

英文版本：英文版本有 4 种，正题名的形式有“The three musketeers”“三个火枪手”，以全一册的形式出版，其中 A17 是英文缩写本，由上海外语教育出版社于 2003 年出版，B16 由 Ladybird 出版社于 2008 年出版，A9、A24 由外语教学与研究出版社分别于 2011 年，1994 年出版，但是 A9 的译者是郝运、王振孙，A24 的其他责任者是程静英。

法文版本：法文版本有 1 种，为 A23。原著经勒马歇尔改写，韩伏秋注释。正题名被译为三个火枪手，于 1991 年由商务印书馆出版。

中英对照本：中英对照本有 3 种，其正题名形式有“三个火枪手”“三剑客”，全部以全一册的形式出版，B19 由中国大百科全书出版社于 2001 年出版，A13 由航空工业出版社于 2007 年出版，A20 由外语教学与研究出版社在 1985 年出版。

中法对照本：中法对照本有 1 种，题名全部被翻译为“三个火枪手”，其中 A8 由 R. de Roussy de Sales 改编，李洪峰翻译，于 2011 年由北京语言大学出版社出版。

精装本：精装本有 5 种，全部是中文译本，翻译者有李玉民、周克希、罗国林、王振孙等，其中 3 种以全两册的形式出版，2 种以全一册的形式出版，其中精装版本多集中在 2013—2015 年出版。

注释本：原著注释本有 3 种，A17 的正文是英文，注释是中文；A24 正文是英文，注释是英文；A23 正文是法文，注释是中文。3 种文献都是以全一册的形式出版，其出版社、出版年和其他责任者均不同。

续本：续本有 6 种，所有续本的正题名是“二十年后”，全部为中文译本，部分图书采用精装的形式出版，其中 3 种图书以全两册的形式出版，另外 3 种图书以全一册的形式出版，出版年从 1982—2014 年。

5.3　结果分析

以法国作家 Dumas Alexandre 的作品 *Les trois mousquetaires* 在东南大学图书馆书目数据中进行版本挖掘，发现该文献版本类型十分丰富，主要有中文版、英文

ISSN 1004-325X | CN23-1331/G2

版、法文版、中英对照版本、中法对照本、续写本、注释本、精装本 8 种类型；涉及 20 个译者和注释者，其中李玉民翻译出版的图书种类最多；文献出版时间跨度从 1978 至 2017 年，约 30 年之久，涉及出版社多达 24 家，其中上海译文出版社、上海三联书店出版的图书较多。这些文献版本聚集信息不仅可为 *Les trois mousquetaires* 作品的文学分析与研究，而且可为图书馆经典图书的筛选和导读工作提供帮助。

当前汇文系统的 OPAC 检索基本无文献版本聚集功能，在 OPAC 系统中需分别以 *Les trois mousquetaires* 不同改版题名进行检索，并加以人工判断才能识别出少量版本信息。若以当前在机读数据中广泛采用的以统一题名的方式进行版本挖掘，能够检索出以《三个火枪手》《三剑客》为正题名的文献，但数量亦十分有限，主要是由于部分文献在著录时未著录统一题名项，同时每篇文献的统一题名的著录不同，使得难以实现对所有版本的聚集。而本文建立的文献版本挖掘模型能够识别出以《三个火枪手》《三剑客》《The three musketeers》《二十年后》为文献正题名的 33 种原著同种文献，能够起到较好的版本挖据与聚集功能。本文研究不足之处在于，对于原著的另外两篇同种文献，即正题名分别为《侠隐记》和《三剑侠》的文献，却没有能够进行有效聚集。

## 6 结　语

本文通过对版本的发现过程，同一种文献和同种文献不同版本的认定，常用的文献版本聚集方法及版本数据在机读数据中的表现，构建了文献版本关系识别与挖掘模型，以 *Les trois mousquetaires* 作品为例，以“三个火枪手”作为检索初始入口，能够实现因装帧不同、出版社不同、出版时间不同、版次不同、语种不同等同种文献的发现。由于本文构建的文献版本关系识别模型具有滚动性，因此以“三剑客”“The three musketeers”“二十年后”或“效忠国王”等为初始检索入口，同样能够达到以“三个火枪手”作为初始检索入口的挖掘效果。另外，因书目元数据本身在著录的过程中一些人为因素存在一定的质量问题，会影响识别和版本聚类过程。在进行同种文献识别过程中，文献 200$a 正题名、5-- 相关题名块是识别因题名不同的同种文献的关键字段，不同题名的文献之间也因 200$a 正题名、5-- 相关题名块之间存在一定的关系，所以能被识别出，若同种文献题名之间没有任何关系，则很难被识别出，例如本文研究中 Dumas Alexandre 的作品 *Les trois mousquetaires* 被翻译为“侠隐记”和“三剑侠”没有作为同一种文献被识别出。因此对于题名变动较大或改换题名的同一种文献的识别将成为本文今后进一步研究的方向。

**参考文献：**

[1] 中国新闻出版广电网 .2016 年新闻出版产业分析报告 ( 摘要版・上 )[EB/OL].[2018-05-20].http://www.chinaxwcb.com/ 2017-07/25/content_358659.htm.
[2] 曾伟忠 . 我国文献编目规则著录信息源条文存在的问题述评 : 以图书为例 [J]. 图书馆学研究 ,2018(4):67-72.
[3] 寿玉清 . 关于图书版本认定及著录的思考 [J]. 情报探索 ,2008(1):108-110.
[4] 李　文 . 对中文图书版本项信息源不一致情况的分析与处理 [J]. 图书馆论坛 ,2009,29(5):110-112.
[5] 王玮琦 . 中文图书版本浅论 [J]. 山东师范大学学报 ( 人文社会科学版 ),2002(4):110-111.
[6] 马兰芳 . 版本确认问题浅谈 – 针对我馆图书著录情况谈版本的确认与著录 [EB/OL].[2018-05-22].http://www.docin.com/p-1207390840.html.
[7] 李庆文 . 论同种书及其复分 [J]. 图书馆学刊 ,2014,36(1):44-47.
[8] 富　平 , 黄俊贵 . 中国文献编目规则 [M].2 版 . 北京 : 北京图书馆出版社 ,2005.
[9] 朱培育 , 马书慧 . 普通图书著录规则图例外手册 [M]. 沈阳 : 辽宁人民出版社 ,1986.
[10] 赵伯兴 . 翻译图书之不同版本的书目关系建构研究 [J]. 国家图书馆学刊 ,2005(2):50-52.
[11] 王玉梅 . 多版本文献书目关系之揭示 [J]. 图书馆理论与实践 ,2007(3):76-78.
[12] 梁美宏 , 曾建勋 . 基于书目关联数据的文献版本关系发现研究 [J]. 图书情报工作 ,2016,60(9):123-130.
[13] 胡小菁 .BIBFRAME 核心类演变分析 [J]. 中国图书馆学报 ,2016,42(3):20-26.
[14] 曹月珍 , 马建玲 . 关联数据在图书馆的最新发展 [J]. 图书馆学研究 ,2014(14):6-12.
[15] 何　云 , 黄久斌 . 中文同种书的版本类型与著录举例 [J]. 现代情报 ,2004(1):166-168.

[ 作者简介 ]
赵娅娜　女，东南大学经济管理学院图书情报与档案管理专业硕士研究生。
常　娥　女，博士，东南大学图书馆副研究馆员，硕士生导师。

[ 收稿日期：2018-07-10]