DOI: 10.13382/j.jemi.2017.04.021

东巴古籍资源的数字化及数据管理*

王玉娇 耿 思 李 宁
(北京信息科技大学 计算机学院 北京 100101)

摘 要:目前大多数东巴经典原始手稿被十多个国家的著名机构收藏,学术研究处于分散形态,沟通不便。构建东巴古籍共享平台有利于经典文化的抢救与传承。针对东巴古籍资源的数字化以及数据存储的问题,在分析现有信息抽取方法以及数据存储方式的基础上,提出了《中国少数民族古籍总目提要(纳西卷)》纸质书籍的数字化方法,并使用元数据表示从纸质书籍中抽取的东巴古籍书目,最终使用XML数据库管理数字化后的内容。实验结果表明,提出的信息抽取方法能够针对东巴古籍书目的特殊结构正确地抽取内容,并提供结构化检索手段。验证了该方法的可行性、正确性。这项研究对于少数民族古籍的数字化以及半结构化数据管理具有重要的借鉴意义。

关键词: 古籍资料数字化;信息抽取;XML数据管理
中图分类号: TP391.1 文献标识码: A **国家标准学科分类代码**: 520.6010

Digitalization and data management of Naxi Dongba manuscripts

Wang Yujiao Geng Si Li Ning
(School of Computer Science, Beijing Information Science & Technology University, Beijing 100101, China)

Abstract: At present, most original classic manuscripts of Dongba script have been collected by well-known institutions from more than ten countries. As academic researchersare decentralized, it is very inconvenient for them to communicate with each other. The construction of a sharing platform for ancient books of Dongba script is beneficial for emergency treatment and inheritance of classic culture. In allusion to digitalization and data storage of ancient book resources of Dongba script, a digitalization method is presented in this paper for printing books known as Annotated General Catalog of Ancient Books of Ethnic Minorities in China (Naxi Volume) based on the analysis of existing information extraction approaches and data storage modes. Moreover, metadata is also adopted to refer to the bibliography of ancient books of Dongba script, which are extracted from printing books. And ultimately, XML database is employed to manage the digitalized contents. According to the experimental results, the information extraction approach proposed in this paper is able to extract contents accurately direct at the elaborate structure of the bibliography for ancient books of Dongba script on one hand and provides structured retrieval means on the other hand. As a result, both feasibility and validity of such an approach areverified. This research has important reference meanings for the digitalization and semi-structured data management of ancient books of ethnic minorities.

Keywords: digitalization of ancient books; XML data management; information extraction

1 引 言

本文的研究工作是以国家社会科学基金重大项目(12&ZD234)——“世界记忆遗产”东巴经典传承体系数字化国际共享平台建设研究课题下的子课题三为研究背景。东巴古籍资源的数字化主要分为编目资源的数字化以及东巴象形文字的数字化[1]。本文的研究工作主要是

收稿日期:2017-01 Received Date: 2017-01
*基金项目:国家社会科学基金重大项目(12&ZD234)、东巴经典古籍基础数字档案建设与设计(KF20161123206)、东巴经典传承体系国际共享平台视频设计(KF20161123207)资助项目

编目资源的数字化,将《中国少数民族古籍总目提要(纳西卷)》(简称《总目提要》)中的内容,包括书中的图片等资料信息,通过数字化的手段,存储到数据库中,并对外提供检索。

作为本文的重要参考,《总目提要》共收录纳西族古籍条目 1834 条,其中甲编为书籍类,收录东巴古籍 1373 条,汉文古籍 16 条;乙编为铭刻类,收录东巴文铭刻 7 条,汉文铭刻 15 条;丙编为文书类,收录东巴文文书 39 条,汉文文书 2 条;丁编为讲唱类,收录 382 条。另附东巴古籍存目 8174 条。书籍类古籍条目参照《纳西东巴古籍释注全集》(100 卷)的分类方法,分为祈福延寿类、禳鬼消灾类、丧葬超度类、占卜类、其他杂类等 5 大类。讲唱类亦按纳西族古籍特点分为神话传说、民间故事、史诗歌谣等 3 类,"讲唱"部分因无文字记载,只用汉文标题。《总目提要》中古籍书目的注音符号采用《纳西文字方案》并提供了《纳西文与汉语拼音和国际音标对照表》。古籍书目正文中的专有名词,按东巴文化研究所编印的《纳西东巴经专有名词汉译规范》统一。

在适用范围方面,基于统计原理的方法适用于以文字为主题并且文字部分相对于其它部分来讲具有明显数量优势的数据,针对不同的数据要应用不同的特征。基于模板的方法适用于相似度较大的文本,如内容特征相似的文本。

在复杂性方面,基于统计的方法在理论上易于实现,但其难点在于确定一个合理的阈值,阈值的确定对信息抽取的准确率产生直接影响。基于模板的方法免去了对同类文本的重复操作,针对相似的文本结构总结出统一的抽取模板,但在模板的生成方法和模板通用性方面还有待改善。

在分析现有信息抽取方法的基础上,选择一个适当的信息抽取方法是本文的重要研究工作。在信息抽取之前,必须对以下 3 个问题进行分析:应依照何种规则进行古籍书目抽取工作;抽取后的内容以何种形式进行表示;数字化后的内容如何进行查询和管理。

2　相关研究

2.1　信息抽取相关技术

在信息抽取领域,存在两大传统的具体实现方法,分别为基于规划的方法和基于统计的方法。

基于规则的抽取方法历史比较悠久,其规则库的质量决定了抽取结果的质量。规则库一般由专业人员进行编制,通过归纳整理出实体及其关系识别的知识并通过一定的格式描述。文献[2]中给出了一种用于互信息的术语抽取系统,其普通词语库就是通过手工方式建立的。手工构建规则库一般具有较高成本,而且应用领域比较单一。因此在很多自动学习规则的信息抽取系统中,多种抽取模型[3-5]被设计出来并可使用多种规则的表达方式[6-8]。其中有两种比较常用的方法:1)使用基于归纳的逻辑规划方法(ILP)进行逻辑规则的学习;2)通过定义模式规则以对实体及其关系进行抽取[5]。基于规则的信息抽取,无论是手工编制还是自动学习,都需要人工参与,对于小规模单一主题的抽取会有比较满意的效果,但在大规模多主题的海量数据中,往往表现出精度差,适应性不强等弊端。

基于统计的抽取方法不需要编制复杂的规则库,处理海量无结构的自由文本具有优势。基于统计的抽取方法起源于序列标注问题,如词性标注系统或语法标识系统使用动词、名词和形容词等通过语义和语境的判别来标记文本内容,并取得较好的效果[9]。隐马尔科夫模型[10-13]和最大熵模型[14]是两种较为成熟的统计模型,在对序列进行标注的同时考虑到输入序列特征之间的相关性,因此在很多应用领域中取得了较好的效果[15-16]。

本文采用基于规则的方法进行古籍信息抽取,选择此方法与《总目提要》中的古籍书目特点有关,由于古籍书目信息的相关词不多,表达形式尽管多样但在特性领域中有限,且构成比较简单,有较为明显的规律性,因此采用基于规则的方法对古籍书目进行抽取会比较有效。又由于正则表达式用于字符串匹配功能强大,得到大多数程序设计语言支持,故本课题采用正则表达式来描述古籍书目的抽取规则。

2.2　东巴古籍元数据规范

文献[17]提出了一种东巴古籍元数据规范。它是以 DC(dublin core)元数据为基础,针对东巴象形文字及其数字资源的特点提出的一套元数据规范。该元数据规范可以描述的东巴资源类型有东巴手稿、扫描件以及释读录音录像等内容。DC 元数据规范的 15 项元素有 12 项在东巴元数据规范中重用。该元数据规范可以基本满足国际共享平台的要求,本文将以此元数据规范中的数据项作为东巴古籍书目的抽取依据。文献[17]提出的元数据结构如图 1 所示。

东巴古籍元数据规范以古籍节点作为根节点,有纸质古籍、扫描古籍、古籍录音和古籍录像 4 个子节点。本文研究工作是将提取完成的《总目提要》中的古籍书目内容存储到纸质古籍子节点中,因此在本文中将着重介绍纸质古籍节点及其子节点。图 1 所示的是纸质古籍节点下的元数据结构。纸质古籍节点下的元素名称、元素修饰词、元素修饰词定义以及复用标准如表 1 所示。

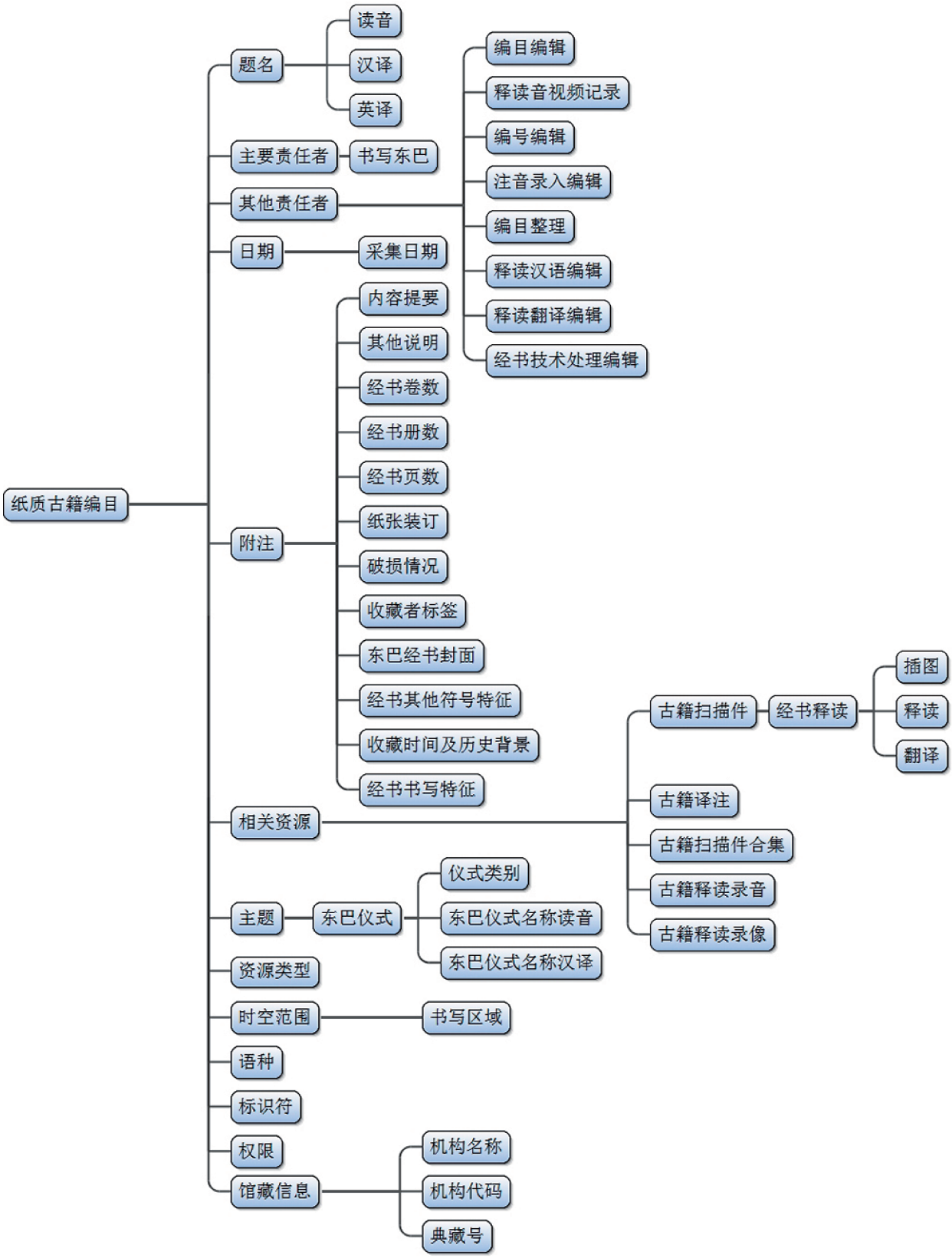

图1　东巴古籍元数据规范 Scheme 结构

Fig. 1　Schematic diagram of Dongba metadata specification Scheme structure

表 1　东巴经古籍元数据规范元素列表

Table 1　Element list for metadata standard of ancient books in Dongba script

名称	元素修饰词	元素修饰词定义	复用标准
题名	东巴经书名称读音	东巴经书的名称读音	dc:title
	东巴经书名称汉译	东巴经书的名称汉文翻译	
主要责任者	书写东巴	编写这部经书的东巴祭司名字	dc:creator
其他责任者	编目编辑	为这部经书做编目录入工作的人员	dc:contributor
日期	采集日期	数字化(照相、扫描)的日期	dc:date
附注	内容提要	经书内容要点	dc:description
	经书卷数	经书的总卷数	
	经书册数	经书的总册数	
	经书页数	经书的总页数	
	经书其他符号特征	经书外部特征的其他描述信息	
	纸张装订	经书的装订情况	
	破损情况	经书的破损情况	
	收藏者标签		
相关资源	古籍扫描件	对古籍数字化处理后的静态图片,这里是对这一静态图片数据的标识符的引用	dc:relation
	古籍释注	无	
	古籍扫描件合集	古籍扫描件是对古籍数字化处理后的静态图片,扫描件合集是对多个扫描件的打包资源,这里是对这一合集数据的标识符的引用	
	古籍释读录音	请东巴对东巴经典古籍进行释读过程的录音资料,这里是对这一资料的标识符的引用	
	古籍释读录像	请东巴对东巴经典古籍进行释读过程的录像记录资料,这里是对这一资料的标识符的引用	
主题	东巴仪式	无	dc:subject
	东巴仪式名称读音	东巴仪式名称的读音	
	东巴仪式名称汉译	东巴仪式名称的汉文翻译	
资源类型		有关资源内容的特征和类型	dc:type
时空范围	书写区域	书写区域	dc:coverage
语种		资源内容的语种	dc:language
标识符	东巴经古籍标识符	东巴经古籍的唯一标识	dc:identifier
权限		针对资源的访问权限	dc:rights
馆藏信息	机构名称	收藏经书的机构名称	mods:location
	机构代码	收藏经书的机构代码	
	典藏号	收藏单位为了检索和排架的需要而给予每一本经书的特定号码	

2.3 《中国少数民族古籍总目提要(纳西卷)》

以《总目提要》中,甲编书籍类的古籍“祭天·远祖回归记”为例,其印刷的内容如图 2 所示。从图 2 可以看出,甲编书籍类的书目可以提取出卷数、册数、页数、作者、分类、纸张装订、其他符号特征、破损情况、收藏、译本、出版等信息项。同理,丙编文书类的书目可以提取出纸数、作者、分类、内容、纸张装订、破损情况、收藏等信息项。丁编讲唱类的书目可以提取出分类、流传地、内容、作者、记录篇幅、收藏等内容。

《总目提要》的正文书题是古籍内容的重要组成部分,一般采用汉文、东巴文、纳西文字的方案排序(图 2),此外,其编目中还包含国际音标。因此要求数字化后的信息表示和管理能够支持东巴象形文字、纳西文字和国际音标。本课题中,采用了 GB13000 (Unicode)编码字符集,同时采用了昆明理工大学开发的东巴象形文字和纳西文字的扩展字库和输入法[18],并采用云龙国际输入法[19]以支持国际音标的输入。

3 《总目提要》的内容抽取方法

3.1 规则模板的设计

总目提要的数字化的过程主要是,首先将纸质书籍《总目提要》扫描成图片存储为 PDF 格式的文件。然后通过 OCR(光字符识别),得到文本文件和图片文件,其

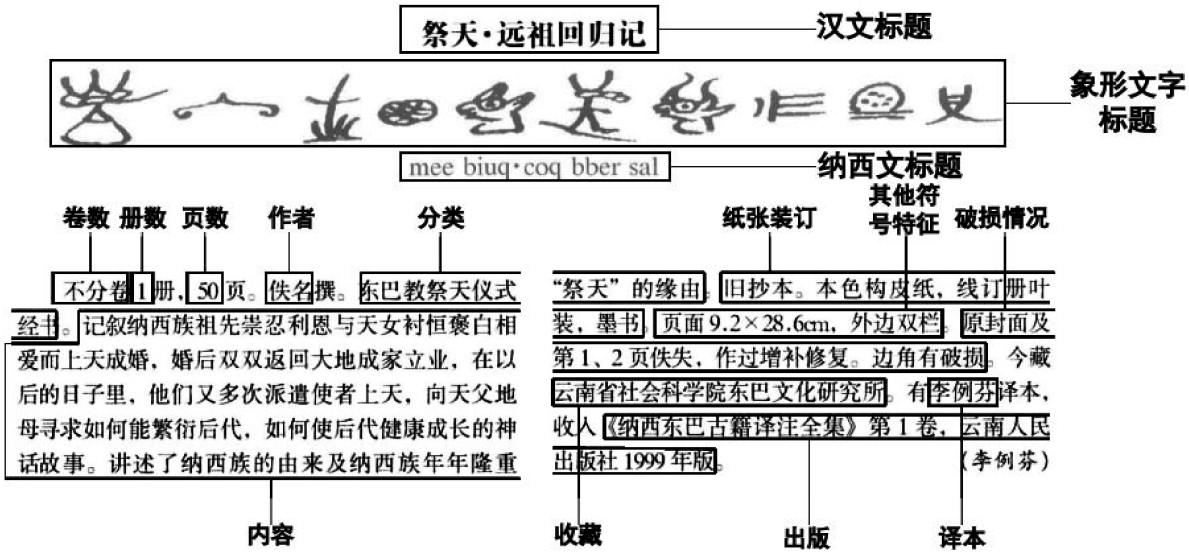

图2　“祭天·远祖回归记”印刷内容

Fig. 2　The printed content of “Sacrifice to Heaven · Return of a remote ancestor”

中文本文件中包含汉文标题、东巴文标题以及古籍书目内容，图片文件中包含象形文字。对提取出的文本文件进行校对，去除错误，最后将文本文件根据模板抽取成半结构化的 XML 文件。对识别出的象形文字图片命名为以汉文标题为名称的图片。具体过程如图 3 所示。

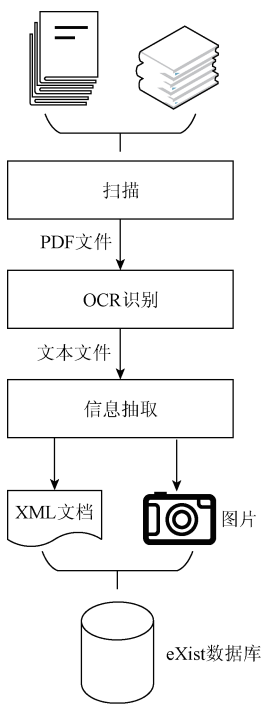

图3　纸质书籍的数字化过程

Fig. 3　Digitation process of printed books

这里重点论述如何对校对后的文本内容进行信息提取。此信息抽取的目标是将《总目提要》文本文件中的内容抽取出来，形成符合东巴古籍元数据规范（图 1）的半结构化数据。

在信息抽取之前，首先要分析古籍书目的文本内容，针对元数据中的不同内容，分析文本的内容特征。本文采取的信息抽取方法是基于正则表达式的模板抽取方法。以甲编中的某篇古籍“祭天·远祖回归记”为例（图 2），可以将其内容归纳成卷数、册数、页数、作者、分类、内容、纸张装订、其他符号特征、破损情况、收藏、译本、出版等内容。根据正则表达式的书写规则，设计针对甲编书籍类古籍书目设计的提取模板如图 4 所示。

```
经书卷数：(((?<=分)|^)([0-9]+)(?=卷)|(不分卷))
经书册数：(((?<=卷)|^)([0-9]+)(?=册))
经书页数：([0-9]+)(?=页)
书写东巴：(((.*?)|^)([\u4e00-\u9fa5])+)(?=撰)
东巴仪式名称汉译：(((.*?)|^)(((.*?)|^)(([\u4e00-\u9fa5])+)经书|画谱|规程|零杂经|用书|占卜书|卜书|卜卦书|歌范本|工具类书|家谱))
内容提要：null
纸张装订：(((.*?)|(.*?)(抄本。(本色|棉纸|线订|墨书).*?)|((抄本|本色|线订|墨书).*?))。
经书其他符号特征：((页面.*?栏.*?)|(页面.*?)|(页[0-9].*?栏.*?))。
破损情况：null
收藏者标签：((?<=今藏|收藏于|现藏|。现存|今存).*?)。
馆藏信息：null
```

图4　甲编书籍信息抽取模板规则

Fig. 4　Template of edition A

以模板中的第一条规则提取“经书卷数”信息为例，在提取“经书卷数”信息时可以分为两种情况：一种是以“分”作为开始（？≤分|^），或者以数字“0～9”作为开始，数字“0～9”匹配以此或者多次（[0～9]+），以“卷”作为结束（？=卷）；另一种情况是“不分卷”（不分卷）。通过此提取模板就会有 3 种类型的卷数信息，分别为：“分 x 卷”，“x 卷”，“不分卷”（x 代表 0～9 中的一个或者多个任意数字）。其他信息项均根据抽取模板规则进行抽取。具体的信息抽取算法如表 2 所示。

表 2　信息抽取算法

Table 2　Algorithm of information extraction process

Algorithm: Information Extraction

Input: regular expressions template ***R***,
the literature manuscripts text ***T***

Output: Well-formed XML object ***X***

1. Initialize XML object ***X***
2. Extract record by matching ***R***
3. Put into ***X***
4. Repeatstep 2 to 3 until the end of ***T***
5. Return ***X***

根据图 4 甲编书籍类的抽取模板对“祭天 · 远祖回归记”进行信息抽取,抽取结果对比如图 5、6 所示,图 5 为信息抽取前的古籍内容,图 6 为信息抽取后的 XML 文档。

祭天 · 远祖回归记
mee biuq · coq bber sal
不分卷1册，50页。
佚名撰。
东巴教祭天仪式经书。
记叙纳西族祖先崇忍利恩与天女衬恒褒白相爱而上天成婚，婚后双双返回大地成家立业，在以后的日子里，他们又多次派遣使者上天，向天父地母寻求如何能繁衍后代，如何使后代健康成长的神话故事。讲述了纳西族的由来及纳西族年年隆重“祭天”的缘由。
旧抄本。本色构皮纸，线订册叶装，墨书。
页面9.2x28.6cm，外边双栏。
原封面及第1、2页佚失，作过增补修复。边角有破损。
今藏云南省社会科学院东巴文化研究所。
有李例芬译本，收入《纳西东巴古籍译注全集》第1卷，云南人民出版社1999年版。(李例芬)

图 5　信息抽取前的文本

Fig. 5　The text file before extraction

3.2　抽取结果

衡量信息抽取的性能主要根据两个评价指标:召回率和准确率。召回率等于系统正确抽取的结果占所有正确结果的比例;准确率等于系统正确抽取的结果占所有抽取结果的比例。为对本文提出的信息抽取方法进行测试,选取了《总目提要》中甲编书籍类古籍 1 373 条作为测试集进行信息抽取,召回率和准确率均达到 100% 。

上述结果说明,本文提出的基于规则的信息抽取方法能够针对东巴古籍书目的特征,正确地抽取古籍内容。

4　《总目提要》的数字内容管理

《总目提要》数字化之后,面临的问题即如何有效地管理和检索数字化内容,关键是数据库管理系统的选择。根据本课题的需求,数据库管理系统应该可以便于书目结构的变化和扩展,适应新入库资料的编目需求。此外,作为共享平台还应支持古籍的结构化查询,以及支持东巴象形文字、纳西文字和国际音标。

目前主流的数据库分为关系型和非关系型数据库,

```
<db:纸质古籍>
  <db:题名>
    <db:读音>mee biuq · coq bber sal</db:读音>
    <db:汉译>祭天 · 远祖回归记</db:汉译>
  </db:题名>
  <db:主要责任者>
    <db:编目编辑/>
  <db:日期>
    <db:采集日期/>
  </db:日期>
  <db:附注>
    <db:内容提要>记叙……的缘由</db:内容提要>
    <db:经书卷数>不分卷</db:经书卷数>
    <db:经书册数>1</db:经书册数>
    <db:经书页数>50</db:经书页数>
    <db:纸张装订>旧抄本。本色构皮纸，线订册叶装，墨书</db:纸张装订>
    <db:破损情况>原封面及第1、2页佚失，作过增补修复。边角有破损</db:破损情况>
    <db:收藏者标签>无</db:收藏者标签>
    <db:经书其他符号特征>页面9.2x28.6cm，外边双栏</db:经书其他符号特征>
  </db:附注>
  <db:相关资源/>
  <db:主题>
    <db:东巴仪式>
      <db:仪式类别>祈福延寿类</db:仪式类别>
      <db:东巴仪式名称读音/>
      <db:东巴仪式名称汉译>东巴教祭天仪式经书</db:东巴仪式名称汉译>
    </db:东巴仪式>
  </db:主题>
  <db:资源类型>东巴经古籍</db:资源类型>
  <db:时空范围>
    <db:书写区域>未知</db:书写区域>
  </db:时空范围>
  <db:语种>东巴文</db:语种>
  <db:标识符/>
  <db:权限>公开</db:权限>
  <db:馆藏信息>
    <db:机构名称>云南省社会科学院东巴文化研究所</db:机构名称>
    <db:机构代码/>
    <db:典藏号/>
  </db:馆藏信息>
</db:纸质古籍>
```

图 6　信息抽取后的 XML 文档

Fig. 6　The XML file after extraction

XML 数据库是一类主要的非关系型数据库。由于本文将《总目提要》的内容存储成符合元数据规范的 XML 形式,而 XML 数据库能够高效存储 XML 数据以及支持 XML 半结构化查询,便于扩展,应该能够更好的满足本课题的需求。目前存在多种 XML 数据库产品,如 eXist、dbXML、BaseX 等。上述软件均提供了基本的 XML 数据管理功能,其中 eXist 是一款开源的原生 XML 数据库管理系统,近年发展较快,无需预定义存储结构、支持标准的 XQuery 查询语言,可以实现自动索引、扩展的全文本搜索以及对 XUpdate 的支持,并且它与现存的 XML 开发工具紧密集成。本文因此采用 eXist 来管理《总目提要》的数字内容。

eXist 中有几个基本概念,包括 collection(文档集)、resource(资源)和 document(文档)。eXist 通过 collection 将一系列相关的 XML 或者二进制文件聚集起来,其作用类似文件夹。每个 collection 都有一个用 URL 表示的名称。一个 collection 存储所有文件的元数据。resource 指的是存储在 collection 中的文件。document 也是 resource 的一种,指的是一个存储在 collection 中的 XML 文件。

eXist 可以有多个 collection,一个 collection 中可以有多个 resource, resource 可以包含多个命名空间,在 resource 中以命名空间为单位存储多条 XML 数据。

东巴古籍书目资料库系统分为两个 collection,分别

是 dongba collection 和 user collection，dongba collection 用来存储东巴古籍数据，user collection 用来存储用户信息。eXist 数据库的内部存储结构如图 7 所示。

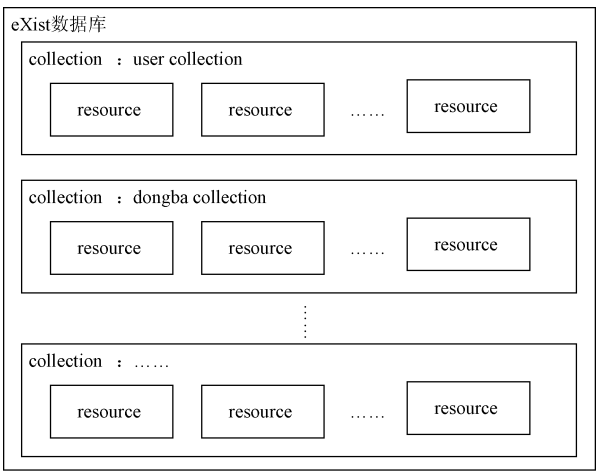

图 7　eXist 的存储结构

Fig. 7　Storage structure of eXist

在东巴古籍资源管理系统中，分为普通用户、系统管理员和数据库管理员 3 种用户角色。普通用户仅拥有检索的权限。系统管理员用户除了拥有检索的权限，还拥有数据添加、修改、管理其上传资源的权限。数据库管理员除了拥有系统管理员的所有权限之外，还可以添加系统管理员用户，管理其下属的所有管理员上传的资源。

5　结　论

本文针对《中国少数民族古籍总目提要(纳西卷)》，提出一种基于正则表达式规则的信息抽取方法。此外提出采用 XML 数据库管理系统对数字化后的内容进行管理和查询，介绍了相关的元数据和数据库设计方法。

实验说明，本文提出的方法可以有效地从文本形式的内容中提取出半结构化的 XML 数据。通过 XML 数据库管理系统可以实现东巴古籍书目的管理和半结构化查询，基本满足本课题对东巴古籍资料的共享和传播需要。在未来工作中，还将对共享平台进行系统设计，并对外提供标准化的检索接口等，以便实现对少数民族古籍书目资源的跨平台检索。

参考文献

[1] 王海燕，王红军，徐小力. 纳西东巴象形文字字符识别研究[J]. 仪器仪表学报，2016，37(1)：61-69.
WANG H Y，WANG H J，XU X L. Study on character recognition of Naxi Dongba hieroglyphs [J]. Instrumentation，2016，37(1)：61-69.

[2] 杜丽萍，李晓戈，周元哲，等. 互信息改进方法在术语抽取中的应用[J]. 计算机应用，2015，35(4)：996-1000，1005.
DU L P，LI X G，ZHOU Y ZH，et al. Application of improved point-wise mutual information in term extraction [J]. Journal of Computer Applications，2015，35(4)：996-1000，1005.

[3] VESELIN S，CLAIRE C. Automatically creating general-purpose opinion summaries from text[C]. In RANLP，2011：202-209.

[4] 黄先珍，杨玉珍，刘培玉. 信息过滤中基于统计与规则的关键词抽取研究[J]. 计算机工程，2012，38(2)：57-59.
HUANG X ZH，YANG Y ZH，LIU P Y. Study of keywords extraction based on statistics and rules in information filtering[J]. Computer Engineering，2012，38(2)：57-59.

[5] 孙荣，周文，刘宗田. 用规则抽取句子中事件信息[J]. 小型微型计算机系统，2011，32(11)：2309-2314.
SUN R，ZHOU W，LIU Z T. Using rules to extract event information from sentences [J]. Journal of Chinese Computer Systems，2011，32(11)：2309-2314.

[6] 蒋德良. 基于规则匹配的突发事件结果信息抽取研究[J]. 计算机工程与设计，2010，31(14)：3294-3297.
JIANG D L. Research on extraction of emergency event information based on rules matching [J]. Computer Engineering and Design，2010，31(14)：3294-3297.

[7] 王吉林，舒江波，李勇，等. 分布式 Web 主题信息抽取的框架探析[J]. 情报理论与实践，2014，37(12)：117-122.
WANG J L，SHU J B，LI Y，et al. Analysis on the framework of information extraction for distributed web topics[J]. Information Studies：Theory& Application，2014，37(12)：117-122.

[8] 冷伏海，白如江，祝清松. 面向科技文献的混合语义信息抽取方法研究[J]. 图书情报工作，2013，57(11)：112-119.
LENG F H，BAI R J，ZHU Q S. A hybrid semantic information extraction methodfor scientific research papers[J]. Library and Information Service，2013，57(11)：112-119.

[9] YAO X. A method of Chinese organization named entities recognition based on statistical word frequency，part of speech and length [C]. Broadband Network and Multimedia Technology (IC-BNMT)，4th IEEE International Conference，2011：637-641.

[10] 梁吉光,田俊华,姜杰. 基于改进 HMM 的文本信息抽取模型[J]. 计算机工程,2011, 37 (20):178-179,182.
LIANG J G, TIAN J H, JIANG J. Text information extraction model based on improved HMM[J]. Computer Engineering, 2011,37 (20):178-179,182.

[11] 李荣,胡志军,郑家恒. 基于遗传算法和隐马尔可夫模型的 Web 信息抽取的改进[J]. 计算机科学,2012, 39(3):196-199,215.
LI R, HU ZH J, ZHENG J H. Improvement of web information extraction based on genetic algorithm and hidden markov model [J]. Computer Science, 2012, 39(3):196-199,215.

[12] 祝伟华,卢熠,刘斌斌. 基于 HMM 的 Web 信息抽取算法的研究与应用[J]. 计算机科学,2010,37(2):203-206.
ZHU W H, LU Y, LIU B B. Improvement of web information extraction algorithm based on HMM [J]. Computer Science, 2010,37(2):203-206.

[13] 宋晓琳,郑亚奇,曹昊天. 基于 HMM-SVM 的驾驶员换道意图辨识研究[J]. 电子测量与仪器学报,2016, 30(1):58-65.
SONG X L, ZHENG Y Q, CAO H T. Research on driver's lane change intention recognition based on HMM and SVM [J]. Journal of Electronic Measurement and Instrumentation, 2016,30(1):58-65.

[14] 孙师尧,妙全兴. 基于改进 HMM 的半结构化文本信息抽取算法研究[J]. 电子科技,2014, 27 (10):111-114,118.
SUN SH Y, MIAO Q X. Algorithm research for semi-structured text information extraction based on hidden Markov model[J]. Electronic Science and Technology, 2014, 27 (10):111-114,118.

[15] 何晓梅. 基于条件随机场的音乐共同语义标注[J]. 电子测量技术,2016,39 (8):70-74.
HE X M. Conditional random fields model for collective annotation of music [J]. Electronic Measurement Technology, 2016, 39 (8):70-74.

[16] 田璟,郭智,黄宇,等. 一种基于多模态主题模型的图像自动标注方法[J]. 国外电子测量技术,2015, 34(5):22-26.
TIAN J, GUO ZH, HUANG Y, et al. Automatic image annotation method based on multi-model topic model[J]. Foreign Electronic Measurement Technology, 2015, 34(5):22-26.

[17] 陈若愚. 针对东巴手稿数字化及共享的元数据规范[C]. 国际测试自动化与仪器仪表学术会议,2014: 330-333.
CHEN R Y. A metadata specification for the digitizing and sharing of Dongba manuscripts [C]. International Symposium on Test Automation and Instrumentation. 2014:330-333.

[18] 张涛,余正涛,郭剑毅,等. 融合特征约束模型的纳西-汉语双语词语对齐算法[J]. 西安交通大学学报, 2011,45(10):48-53.
ZHANG T, YU ZH T, GUO J Y, et al. An alignment algorithm for the Chinese and English words in the fusion model with feature constraint [J]. Journal of Xi'an Jiaotong University, 2011,45(10):48-53.

[19] 李龙,王奕桦. Unicode 国际音标输入法简述[J]. 民族语文,2012 (5):62-64.
LI L, WANG Y H. A brief introduction to Unicode international phonetic input method [J]. Minority Languages of China, 2012 (5):62-64.

作者简介

王玉娇,1992 年出生,北京信息科技大学在读研究生。目前主要研究方向为文档信息处理。
E-mail:baobeijiaohenxingf@ qq. com

Wang Yujiao wasborn in 1992, M. Sc. candidate of BeijingInformation Science and Technology University. The main research field is document information processing.

耿思,1991 年出生,北京信息科技大学在读研究生。目前主要研究方向为文档信息处理。
E-mail: gengsi123@ 163. com

Geng Si was born in 1991, M. Sc. candidate of Beijing Information Science and Technology University. The main research field is document information processing.

李宁,1964 年出生,博士,现任北京信息科技大学教授,目前主要研究方向文档信息处理、XML 及标准化信息技术。
E-mail: ningli. ok@ 163. com

Li Ning was born in 1964, Ph. D., professor of Beijing Information Science and Technology University. The main research field is document information processing, XML and standardization of information technology.