Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

数字图书馆数字化文献再处理工具的开发与实践*

□ 曾文 徐硕 张运良 / 中国科学技术信息研究所 北京 100038

摘要：数字图书馆运用计算机系统管理各种载体文献的加工与服务，通过网络和通信技术支持用户访问数字化文献信息资源。数字图书馆对海量数据的处理能力是保证数据质量、支持与深化数字图书馆服务功能的基础。文章论述数字图书馆中数字文献再处理工具研究的重要性，介绍和阐述已有工作的开展情况，以及结构化的数字文献再处理工具的开发与实践工作。

关键词：数字图书馆，结构化数据，数字化文献，再处理工具
DOI：10.3772/j.issn.1673—2286.2013.07.010

1 引言

21世纪以来，计算机通信与网络技术的飞速发展，使网络已经成为人们获取信息的重要途径，而网络信息资源正在以惊人的速度不断增加，需要存储和传播的信息量也越来越大，信息的种类和形式也越来越丰富，数字图书馆应运而生。数字图书馆作为数字化信息仓储，能够存储大量各种形式的信息，特别是文献信息数据是数字图书馆处理的重点内容之一。即数字化文献数据的处理工作是支持数字图书馆技术服务的数据基础，良好的数字化存储资源是构建优质技术和服务的重要保障。目前，随着各类数字化文献数据资源的丰富，这些来自不同渠道的原始数据格式和质量不尽相同，而且数据量的规模日益庞大。因此，这些数据通过数字图书馆这个窗口面向用户服务之前，必须经过对其进行数字化的一系列基本处理过程。显而易见，自动化处理这些海量的数字化文献数据是必要的。本文的研究工作正是基于这样的研究背景提出和开展的。

2 国内图书馆数字化文献处理的现状

国内数字图书馆经过十几年来的发展，文献资源的数字化建设得到了极大的进步[1,2]。目前多数的图书馆基本上是采用本地加工和外包加工的方式，使用的数字化加工系统有TPI、TBS、TRS、DIPS等数字资源加工系统，这些系统实现已有和现有的纸质文献的基本加工和处理过程，将文献资源制作成为数字化文献信息资源，进行储存和管理，丰富虚拟图书馆的文献信息资源，来进行网络化检索和阅读等服务，从而促进数字图书馆的发展。这类信息资源又可分为结构化和非结构化的数字资源。这些加工和处理实现文档扫描、条目著录、文本化、标引、挂接等一系列操作过程，但实际上这些操作只是文献资源层处理的基本环节，并未实现对数字化资源的深层次数据处理、组织和整合。随着文献资源逐年海量式的递增，这种资源处理方式已经难以更好提高数字图书馆文献的检全率以及检准率，也难以满足图书馆及情报研究机构对文献资源深层次信息挖掘和分析的需要，对这些文献数据资源的再处理和整合技术研究是必要的。

目前，国内外很多研究人员从事对文本信息挖掘和处理的研究工作，并取得相应的研究成果，其中包括对文本数据的关键术语抽取，文本数据内容自动分析、语义分析等涉及数据内容挖掘方面的研究工作[3]。但这些研究成果基本是建立在数据规整、数量规模有限的文本数据实验基础之上的，当实际应用

* 本文受国家自然科学基金项目“支持面向特定情报分析应用的知识组织系统快速构建关键问题研究”（编号：71203208）、“十二五”国家科技支撑计划课题“基于多源信息的电动汽车数据挖掘关键技术研究”（编号：2013BAG06B01）的支持。

于数字图书馆这种海量数据时，无法完全实施已有技术。因为技术的实施是建立在良好的数据之上的，目前数字图书馆的海量数字化文本数据事实上还不具备这种良好的数据质量，如何处理已有和未来的数字化文献资源使之符合技术研究的应用需求，是目前数字图书馆以及情报学研究人员在实际工作中面临和需要解决的主要问题之一。

3 我们的研究工作

3.1 数据分析

近年来，数字图书馆的应用已不仅仅是作为数字文献数据资源的简单原文传递的服务窗口，数字图书馆数据资源的丰富和增加，对于数字化文献的处理、存储、维护和面向用户的检索机制都提出了前所未有的挑战。如何挖掘海量文献数据背后的隐含知识和技术信息、文献之间关联信息[4]，以及学科技术研究趋势分析和预测等，都是图书馆及情报学研究领域开展研究的技术热点。但是开展这些研究面临的首要问题都是数据的获取和处理问题，已有的数字化文献加工处理方式并不能满足这些需求。此外，通过我们对数字图书馆现有的数字化资源进行实际调研发现，这些已加工处理的数字化数据资源的质量和规范程度，距离现有技术的实用化实现还有很大的差距。主要表现在如下几个方面：

（1）数据的存储内容存在加工或录入的错误，这些错误的存在对于海量数据集来说，人工识别和解决都是相当困难的，智能化加工处理技术是必须的。

（2）国内不同的加工单位或文献供应商由于采用数字化加工方式不同导致数据存储的结构、描述等不尽相同，数据需要进行结构映射和结构描述归一化加工。

（3）对于购买的国外数据库的数据，我们分析时需要从数据库中导出相应的数据，这些数据导出后的格式同样存在需要二次格式转换和加工的问题。

（4）对于数据内容的深层次信息挖掘和分析需要涉及更多数据内容的细节，不单单是目前已加工的文章标题、摘要信息等数据字段，还要涉及如中文作者姓名消歧、外文作者姓名要区分作者的姓与名的信息，作者单位消歧、引文、正文等数据信息。对于这些特殊数据字段的内容，现有的数字资源数据库基本并未提供直接可用的内容及文本格式，所以需要对已有数字化数据进行智能化的再处理，人工再处理是不现实的。

基于上述数据分析的情况，开展对已有数字化文献再处理工具的开发与实践探索是必要的。

3.2 研究工作的意义

对于数字图书馆的数字化文献资源进行再处理的重要意义在于，一是对海量数据信息的深层次挖掘技术的实施需要数字化文献资源再处理过程来提高现有数据的质量。二是数字图书馆目前提供给用户的查询检索服务需要改变目前单纯依赖加工的题录数据中作者的标题、关键词和摘要信息进行检索、简单的推送原文的展示数字图书馆的数字化文献数据的方式，这种推送和展示方式使得数字图书馆的服务单一化，缺乏深度知识的推介功能，不符合用户对数据信息的深层次技术信息需求的需要。尽管很多研究机构已经在从事这些方面的研究工作，但是研究常常是独立的，并源自局部的、数量有限的数据来从事研究工作，即这些数据来源并非完全取自数字图书馆的真实数据而做的研究工作，所以其应用性欠缺。而在图书馆研究领域，对于数据再处理研究工作，往往是基于需要去抽取已有数据库的数据，进行实验研究，并未形成实用化处理工具。因此，对于数字化资源的再处理进行实际的开发与实践工作是必要的。我们的研究工作首先是基于现有数字图书馆中的结构化数据资源，开展相应的研究和实践工作。

4 国内数字图书馆数字化文献的再处理

国内数字图书馆目前除了具有中文文献数字化资源外，还包括外文文献数字化资源，其中对部分外文文献数据的结构化处理方式与中文文献一样，也是通过扫描、条目著录、文本化、标引、挂接等一系列基本操作过程，其他外文文献则是购买的全文数据库，通过链接访问国外文献服务机构提供的外文文献资源。对于国内数字图书馆的数字化文献数据，包括结构化数据和非结构化数据，我们对其的再处理直接取自经过一次加工处理后的结构化数字文献数据，进行相应的二次需求处理。目前我们的工作以期刊文献数据为研究重点，对来自不同供应商的结构化数字资源，我们需要统一结构和抽取字段内容重新处理并存储，以为深层次的研究服务，这种深层次研究包括文献之间的内容关联、技术关联、知识信息

挖掘和分析等内容。而对于非结构化的数据处理研究是我们日后的工作重点内容之一。目前，我们已经先后研究并处理了部分结构化数据，如期刊文献数据、专利文献数据和外文数据库数据等。

4.1 结构化数字文献再处理的关键技术和基本处理流程

结构化数字文献再处理涉及的主要关键技术问题是数据的加工和存储技术。首先我们将结构化的数字文献数据导出成可再处理的统一数据格式，例如XML格式。

具体的加工技术包括：1）数据元素的识别，即自动识别数据资源中说明和携带的数字化文献数据资源的信息，重点是对原有结构化数据中并未提供的数据元素信息进行整合和抽取。2）数据内容的清洗，针对结构化数字文献数据存在前期加工处理的错误现象，在数据资源存储之前，首先需要对数据资源进行必要的自动“清洗”处理，去除不规范的字符和符号等，否则导入数据库的过程中会出现不必要的数据导入错误，而且影响日后数据整合和分析质量。

存储技术包括：1）建立数据库，用于存储处理后的数据，实现对数据的修正和消歧结果进行实时存储。2）将自动识别的数据资源内容与存储的数据库中的字段实现自动匹配，并自动存储在相应的数据库字段内。

为此，我们设计了如图1所示的数字化文献再处理的基本处理流程。

流程图中的关键技术环节即实现对数字化数据资源的数据加工和存储，它主要包含：一是数据元素的识别，数据内容的“清洗”处理环节；对于原有结构化数据中已有的数据字段，通过辨识数据字段信息，抽取相应数据字段中的数据内容；对于结构化数据中未加工的数据字段，则需根据整个的数据内容，甚至通过全文数据和网上其他相关资源的内容作参考，设计相应的自动处理方案实现数据整合和抽取；二是实现数据元素与用户的数据库字段名称的自动映射与匹配，并完成对加工处理后的数据内容自动导入用户数据库的处理过程，其中数据库的结构设计要先期设计并完成；三是实现数据的消岐技术，这部分是技术的难点问题，我们也正在探索和实践阶段；四是建立相应的数据处理规范，我们根据当前数据分析和研究的需求，制定相应的数据规范和要求。数据规范是一项长期积累的工作，我们将随着研究和实践工作的推进，逐步完善，形成适用于数字化文献再处理的数据规范和标准。

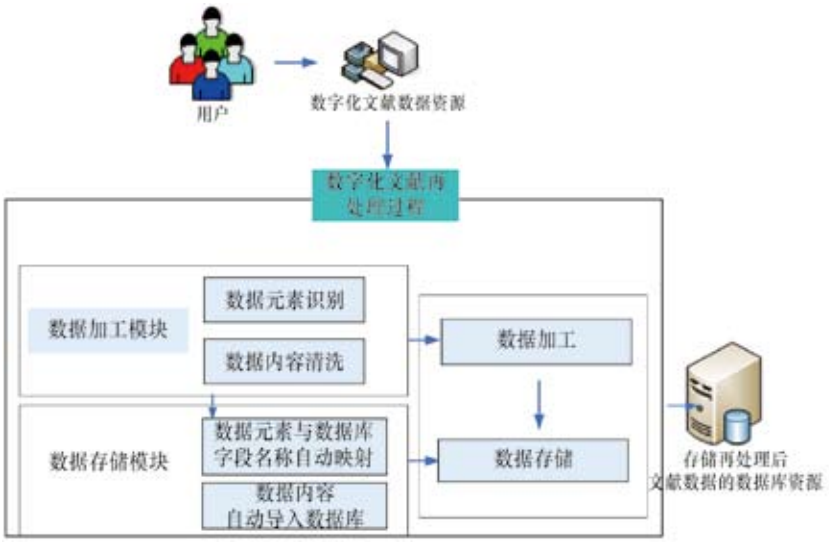

图1 数字化文献再处理的基本流程

基于以上基本处理流程，我们开发了针对数字图书馆的结构化数字文献再处理工具，该工具可以提高数据再处理的效率，满足深层次数据挖掘和分析等研究工作的需要，该工具可以自动实现如下操作过程：

（1）用户提交操作请求，输入待处理的数字化文献数据资源在用户计算机中的存储地址，之后进入数据加工与存储处理过程；

（2）再处理工具自动定位用户输入的存储数据位置，提示用户输入需要加工的数据元素名称，之后再处理工具对数据进行主要数据元素和非主要数据元素的自动识别；

（3）再处理工具对识别出的数据元素对应的数据内容，进行必要的数据内容清洗，例如，自动“清洗”数据内容中首尾出现的不规范字符，并在操作界面上显示识别出所有数据名称；

（4）用户根据再处理工具界面提示内容，输入用户需要存储的数据名称，以及用户用于存储这些数据的数据库信息，例如数据库名称、用户及密码、数据库字段名等；

（5）再处理工具根据用户的输入信息，自动实现数据名称与数据库字段名之间的自动映射和匹配；

（6）再处理工具自动实现对数据内容的抽取，并根据用户需求自动导入数据库中对应的数据表内存储。

4.2 数字化文献处理工具的实现

根据前文的数据分析和再处理流程设计方案，我们开发了数字化文献再处理工具，开发编程语言采用Java语言，JDK1.6.0及以上版本。对硬件设备和系统要求是计算机CPU2.5GHz及以上，内存2GB及以上，至少10G硬盘空闲空间；操作系统支持Windows XP、Windows Server 2000及以上版本，Linux、Unix、MacOS等系统；再处理工具的使用界面图示见图2和图3。目前该工具可以实现对数字化科技文献再处理的基本处理过程，随着研究工作的开展还有待于我们进一步完善。

图4和图5显示的是经过再处理工具处理的数字化文献数据资源最终完成之后的数据存储状态。图示中，我们处理了557个xml格式的文件，数据大小为11.2GB，通过我们开发的再处理工具的自动处理，成功完成加工和存储处理过程，并且按用户需求存放在数据库的不同类别数据表的字段内，最终处理结果是每个表的记录数均为2,781,881条。

5 结语

实现对海量的数字化文献数据资源的再处理，满足数字图书馆的工作人员，以及数字图书馆领域的科研人员对数字化文献数据资源的信息挖掘研究进行数据整合的

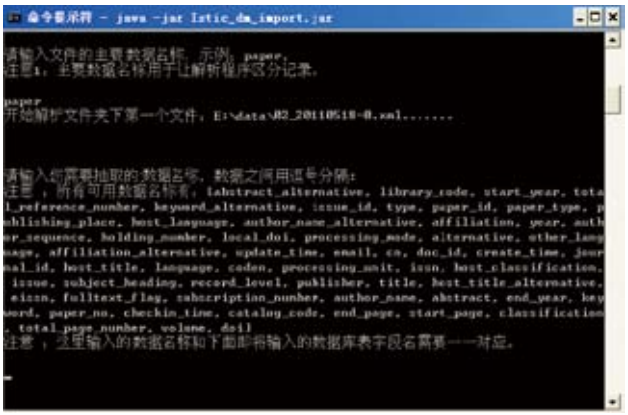

图2 再处理工具的使用界面图示1

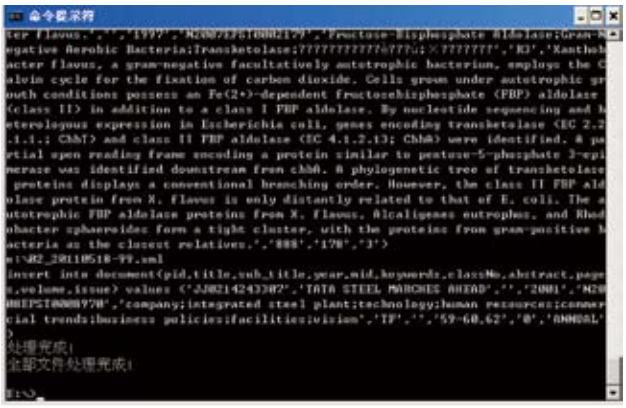

图3 再处理工具的使用界面图示2

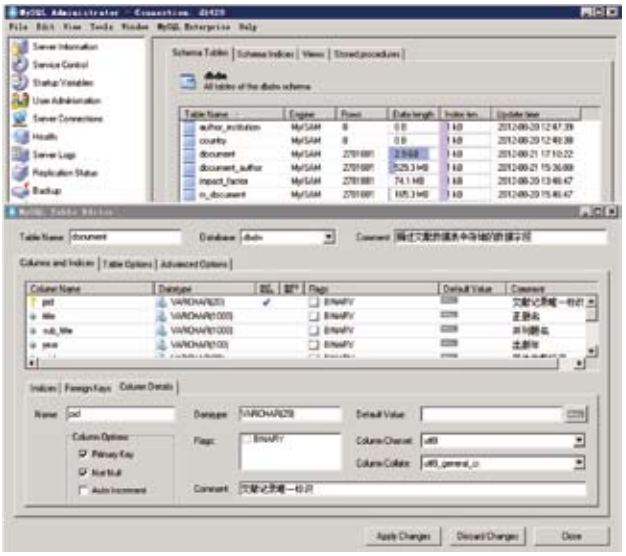

图4 处理后数据库存储状态示例1

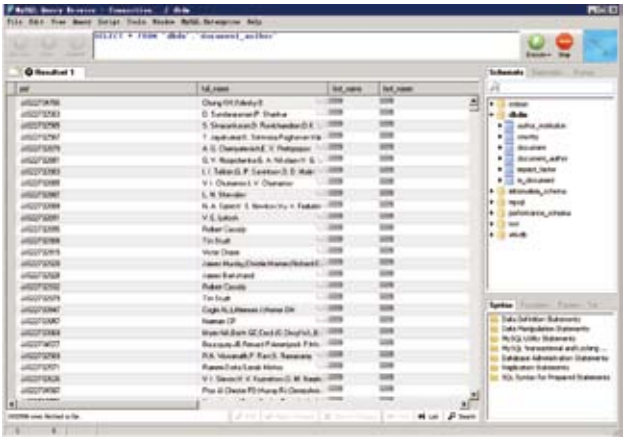

图5 处理后数据库存储状态示例2

需要，是我们研究工作的出发点。目前，我们的研究工作成果已应用于“十二五”国家科技支撑计划课题和国家自然科学基金项目中，并支持这些课题和项目的进一步研究工作。这种再处理工具基本适用于对国家工程技术图书馆和国家科技图书文献中心存储的结构化数字文献数据资源。而对于购买的国外文献全文数据库，数据处理则相对复杂，原因是，国外数字图书馆提供的是检索服务接口，我们对于文献数据的相关信息如关键词、摘要和全文等数据获取需要额外的付费服务。对于这类数字文献，以及非结构化数字资源的处理和研究工作，我们将在未来的研究工作中逐步开展。

参考文献

[1] 赵继海.数字图书馆发展若干领域的评析[J].图书情报工作,200l(3):16-19.
[2] 凌秀丽.略论数字化图书馆与现代化服务[J].图书馆学刊,2005(1):59-60.
[3] THOMAS L C. The State of Mobile in Libraries 2012 [EB/OL]. [2012-07-03]. http://www.thedigitalshift.corn/2012/02/mobile/the-state-of-mobile-in-libraries.2012/.
[4] 林海青,楼向英,夏翠娟.图书馆关联数据:机会与挑战[J].中国图书馆学报,2012,38(197):58-68.

作者简介

曾文，博士，中国科学技术信息研究所，研究方向：智能信息处理、数字图书馆等。E-mail: zengw@istic.ac.cn; zengwen_@sohu.com

The Development and Practice of Digital Library about Structured Digital Document Reprocessing Tools

Zeng Wen, Xu Shuo, Zhang Yunliang / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Digital library uses computer system to manage all kinds of documents processing and service, through the network and communication technology it supports user to access digital literature information. Processing ability of digital library on the mass data is the foundation of ensuring data quality, supporting and deepening the service function of digital library. The paper discusses the importance of data reprocessing tools research, and it introduces the previous work, elaborates the development and practice work of structured digital document reprocessing tools.

Keywords: Digital library, Structured data, Digital document, Reprocessing tools

（收稿日期：2013-01-25）