

【数据库建设】

基于 TPI 的新闻媒体信息数据库建设

——以广州大学图书馆《媒体眼中的广州》全文数据库为例

艾新革(广州大学图书馆,广东 广州 510006/中山大学资讯管理系,广东 广州 510275)

[摘要]本文介绍了新闻媒体信息数据库的发展现状,分析了数据库的基本特点,从构建数据库的数据采集、数据加工、数据发布和输出、数据检索几个环节出发,以广州大学图书馆《媒体眼中的广州》数据库为例,着重阐述了基于 TPI 的新闻媒体信息数据库建设及其建设中注意的问题。

[关键词]新闻媒体;媒体信息;数据库;TPI;数据库建设

[中图分类号]G250.74 **[文献标识码]**B **[文章编号]**1005-6610(2009)08-0025-04

The Building of News Information Database Based on TPI

——Take the Guangzhou News Information Database
of the Library of Guangzhou University as Example

Ai Xin'ge

(Library of Guangzhou University, Guangzhou 510006, China/ Department of Information
Management, Sun Yat-Sen University, Guangzhou 510275, China)

[Abstract] The paper introduces the present situation of news information database development, analyzes the characteristic of database from the processes of building database: data capture, data elaboration, data promulgation and output, data searches. Taking the Guangzhou News Information Database of the library of Guangzhou University as example, it puts an emphasis on the building and many problems of news information database based on TPI.

[Keywords] News; News information; Database; TPI; Building of database

1 概述

新闻媒体是信息传播的重要载体,是报道社会经济、政治文化的主要渠道。随着计算机和网络技术飞速发展,新闻媒体信息网络传播已成为新视点,各新闻媒体注重媒体信息的电子化,相继推出网络版,这不仅方便了媒体信息的网络传播,加速其信息的传播力度,同时也为建设新闻媒体数据库提供了方便,促进了新闻媒体信息数据库的建设与发展。

综观国内外新闻媒体信息数据库建设情况,国外起源较早,始于20世纪70年代^[1],经过三十多年的发展,出现了一大批包括 Factiva、ProQuest、LexisNexis、慧科等新闻媒体信息数据库服务商和服务产品,逐步形成新闻媒体信息开发与服务的产业化,以 Factiva 为例,它将 Dow Jones Interactive 和 Reuters Business Briefing 两大资源库的9000多种新闻媒体信息整合在一起,提供118个国家22种语言出版的重要媒体信息^[2],影响广泛。我国新闻媒体信息数据库起步相对较晚,始于20世纪80年代^[3],且规模较小,档次较低,没有形成产业化,当前最具商业化的新闻媒体信息数据库有三家,分别是:

由新华社和北京电讯工程学院合作开发的《新华社多媒体数据库》;TRS与人民日报社合作开发的《媒体全文按词检索数据库》以及CNKI的《中国重要报纸全文数据库》。就图书馆而言,开发建设媒体数据库以广东中山图书馆开发的《决策内参》系列产品、广州大学图书馆的《媒体眼中的广州》(以下简称《媒体广州》)全文数据库以及温州市图书馆的《媒体看温州》数据库为最成功。

新闻媒体信息数据库除与一般数据库具有共同的技术要求外,还有自身的特点:首先是新闻媒体数据库信息采集量大,仅就报纸而言,全国各类报纸2005年有1926种^[4],其承载的信息当以海量计算;其次是数据库数据源对象信息时效性强,大量信息需及时采集、加工、发布甚至打印呈送到服务对象手中,信息价值时效性很强,时间观念要求较高;第三是采集信息呈现专题性,新闻媒体海量信息决定了搞综合性数据库难度很大,开发建设专题性数据库是比较切实可行的选择,针对用户需求不同,确定数据库信息采集范围,有针对性地加工,最终提供满足客户需要的信息产品;最后是信息内容具有时事性,新闻媒体信息一般是对当前政治法律、社会经济以及科教文卫体等领域的新闻报道,追求快速、准确与简洁,反映的是当前社会发生的一些热点事情,具有时事性特征。

2 基于TPI的新闻媒体信息数据库建设实现思路

2.1 TPI^[5]简介

TPI是由清华同方光盘股份有限公司开发的数字图书馆建设与管理平台——“清华同方数字图书馆管理与建设平台”,是一个全面系统的数字图书馆建设与管理平台软件,是清华同方光盘股份有限公司在建设和管理的知识信息资源库(包括CNKI专业知识仓库和CNKI数字图书馆)的基础上,结合自身开发与应用经验,推出的一套成熟的数字图书馆建设与管理平台。

TPI是基于非结构化文档管理而开发的大型智能内容管理系统,该系统以全文检索数据库(FTS)为核心,采用流行的B/S浏览器的检索方式和先进的三层C/S架构,能够同时管理文字、图片、多媒体等信息,并提供全文检索服务,支持网页的动态发布,是一个面向内容管理的应用、管理和信息发布工具。

TPI系统具有以下突出特点:

- 全文检索基于分词策略,提供中英文混合检索、渐进检索,支持SDK二次开发。
- 提供灵活的内容发布平台,可依用户需要的形式将数据发布到Internet上。
- 提供异构统一检索平台,在统一的检索界面中,可以同时检索多个异构的数据库。
- 提供自动关联功能,用户可以指定库与库之间、记录与库之间、记录与记录之间的关联。
- 提供智能化的电子图书加工工具,具有自动倾斜校正、自动噪声去除、自动二值化、灰度图像页自动搜索与智能二值化等自动图像处理功能。
- 提供了订阅推送功能,每个用户都有自己的特定需求,系统根据用户的需求过滤信息,主动发送用户需要的信息。
- 支持多种国际标准: Dublin Core、MARC、RDF、Z39.50等。

2.2 基于TPI的新闻媒体信息数据库建设工作流程

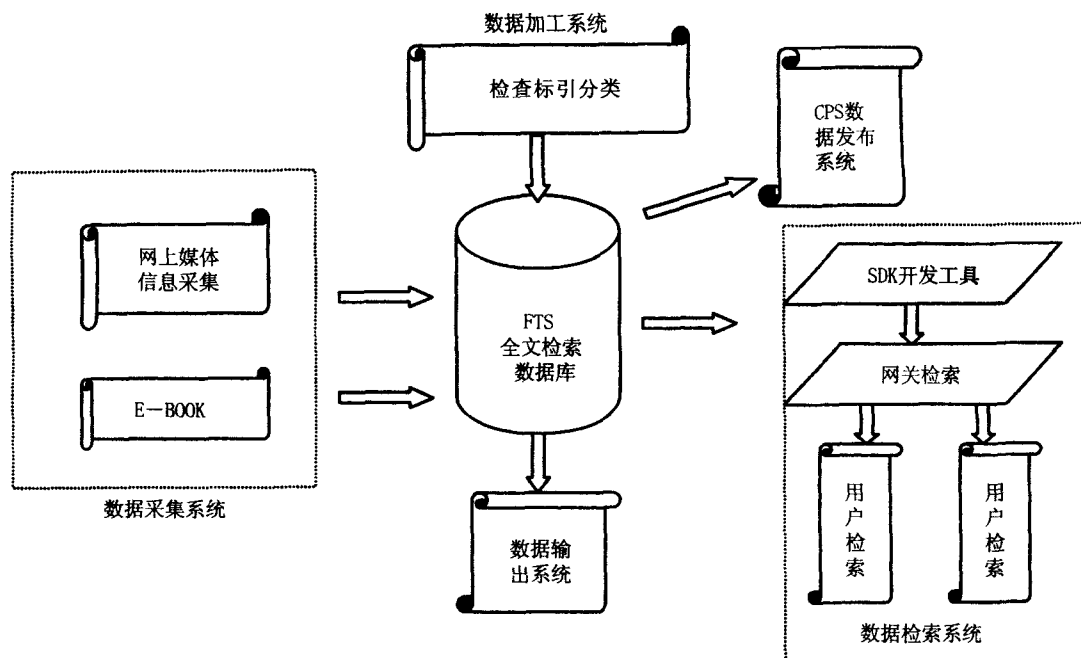
基于TPI构建新闻媒体信息数据库包括数据采集系统、数据加工系统、数据发布和输出系统以及数据检索系统等,要求实现数字资源加工和采集、数据库发布与检索、数字资源管理以及资源数字化、存取网络化和分布化的一整套流程(如图所示)。

(1) 数据采集

TPI数据采集分为两种形式:一是网上媒体信息采集。利用新闻媒体电子版和网上媒体信息发布平台,采集确定主题的媒体信息;二是E-BOOK系统采集。一些媒体信息无电子文档,需要进行电子化处理,E-BOOK系统可以把报刊、书籍、公文等各类纸张文档,通过扫描、图像处理、灰度二值叠加工、目录树加工等过程简单快速地形成电子文档,它支持WORD、PDF、HTML、PS系列、S系统、TXT、PPT等文档格式的转换,支持PDF格式输出。电子文档采集成功后,“采集员”按数据库著录格式进行每个字段的录入,字段包括题名、作者、信息来源、发布日期、版次信息、文摘、备注等,在实际操作过程中,可依数据库的功能要求确定著录字段,其著录格式以图书著录规则为准。因新闻媒体信息数据库字段相对简单,一些著录可自行约定,如《媒体广州》只著录题名、信息来源、发布日期、版次信息等,在对题名著录时,因系统无副题名字段,故凡有副题名者,直接将副题名著录在主题名之后,并用破折号隔开。

在数据采集时,因各种自动采集软件或多或少存在着缺陷,在采全与采准兼顾的情况下,实际操作性较差。虽然人工采集劳动量大,但相比机器采集,人工采集可靠性高,采全与采准率都有一定的保障,实际操作也比较可行。当然,人工采集存在主观性因素干扰,受“采集员”综合素质所制约。

就整个建库流程而言,数据采集是媒体数据库建设中的一个重要环节,它起着数据质量控制作用,其采全率和采准率是评价数据库建设质量的最主要指标。



工作流程结构示意图

(2) 数据加工

一条采集的数据要经过数据加工,才可以正式进入全文检索服务器(FTS)进行发布,数据加工包括三个方面的内容:数据检查、数据标引和数据分类。

数据检查是指对提交数据的真实和准确性进行质量控制,这一操作在数据加工系统中实现。“采集员”采集信息后,进入预提交过程,由“检查员”负责对采集的信息进行筛选,因概念理解偏差,采集时一些无用或相关度不高的信息往往也被采集进来,需要进行审核。另外,一些提交的数据还会出现信息著录的错误,无论是网上 HTML 格式文档录入,还是 E-BOOK 工具转换文档,都可能存在数据失真现象。“检查员”需对出现遗漏和错误的地方进行补漏和修正。

一条记录一旦通过检查,则将该条记录标记为“记录正确”,此记录被锁定,处于不可编辑状态;反之,若记录有误,则下次登录后会看到该记录提示的错误信息,直到修改后标为“记录正确”为止。

TPI 数据标引支持 WORD 文档、NH 文档、HTML 文件、KDH 文件、TXT 文件等文件类型的标引工作,采用可视化操作,直接从原文中选择需要的内容。数据标引存在两种方式:一是通过原文进行标引,对各文件类型记录通过在原文中选取相应文字进行标引;二是通过编辑记录进行标引,对于不能通过原文进行标引的记录,可选择“标引记录/编辑记录”菜单,直接在“记录编辑”对话框中编辑字段即可。

数据分类是利用可视化的操作界面实现数据分类的过程。TPI 数据分类工具提供有标准的《中图法》分类体系,也可采用自定义分类体系,如《媒体广州》将信息分为政治政法、社会、经济、城建、科教文卫体和其它等六大类,每一个大类下再设多级子目,这种分类体系取决于用户单位的实际需求。一般而言,新闻媒体信息数据库以其新闻报道特有的形式和用户群的需求类型,决定其数据分类时无需采用《中图法》等大型分类体系标准,甚至可不进行分类,“以 TPI 强大的检索功能,完全可不采用分类号检索”^[6]。

TPI 数据加工的工具是通过 TCP/IP 协议连接到 FTS 的,在对分配任务进行操作时,数据加工与 FTS 可以不在同一台计算机上运行,满足远程工作和多机运行模式,从而大大提高了数据加工的效率。

(3) 数据发布与输出

数据加工完成后,可以利用 TPI 的“内容发布工具”的“WEB 发布向导”功能进行网上发布,TPI 的内容发布系统(CPS)提供两种发布模式,包括通过客户端管理工具发布和通过 WEB 发布。两种发布模式都能自动完成,无需人工干预,即发即得,立即生效。

在数据发布时,可选择不同的发布模板,CPS 提供包括 CNKI、GOOGLE、FI、OCLC 和图片等不同的风格模式,《媒体广州》采用 CNKI 期刊风格模式。

CPS 功能强大,支持角色管理和用户管理,提供屏幕设计工具,建立多种导航树以及数据库记录之间关联、跳转。为体现人性化设置,支持自定义发布模式,可进行字段的发布选择设置、字段颜色与字体设置和字段的访问权限设置等。在字段的发布选择设置中,可详细设定检索字段、细览字段和概览字段等。

为满足数据输出需要,TPI 有完善的输出系统,可按用户需求进行整体或部分甚至是专题内容的输出。

(4) 数据检索

数据建库的一个主要目的是实现网上共享,使读者可以检索和利用,TPI 采用 IR(Information Retrieval)技术和元搜索(Meta-search)技术,具有全文检索和分布式检索功能。

全文检索采用基于分词的策略,可以同时的词和非词进行检索,其功能主要包括数据库单库检索、跨库检索、视图检索、二次检索、高级检索等检索方式,提供 and、or、not 逻辑操作。在检索项上,分别有题名、关键词、作者、信息来源、日期以及文摘等,各库检索项的多少由数据发布风格以及发布设置来决定。

分布式检索是把分布在不同地理位置的独立自主的多个 TPI 数据库服务器联结成为一个集群系统,这个集群系统中的数据库在逻辑上是一个数据库,对用户是透明的。分布式检索提供跨服务器、跨平台的分布检索形式,用户通过该系统可以最大限度地共享整个集群数据库中的信息,实现分布式、多层次、多类型、特色性的资源共享。

3 新闻媒体信息数据库建设的几点思考

3.1 进行可行性论证,避免仓促建库

图书馆建设新闻媒体信息数据库,因牵涉面较大,投入也较高,在建库前,要基于本馆性质、服务对象、社会责任、用户需求等因素进行认真分析,反复论证,“不能图新鲜,不能拍脑袋,不能盲目,不能攀比,不能一哄而上”^[7],要遵循“用户至上,需求第一”的原则,有目的、有方向、有选择性地建库。最好是采用合作意向性建库,所谓合作意向性建库,就是要先谈定部分用户单位,在具有经费支持的情况下才进行媒体数据库的建设。在具体操作上,可借鉴广州大学图书馆建库模式,广州大学图书馆与广州市市委宣传办合作,先达成部分用户的使用意向,然后启动建库方案。

建设新闻媒体信息数据库工程浩大,在资源采集、数据整理、深层次服务开发以及数据发布与维护等方面需要大量人力、物力和财力,要求图书馆启动建库前要评估本馆的人员配备、技术力量以及财力情况,在条件允许的情况下制定可行性方案。如广州大学图书馆开发《媒体广州》全年配备 10 多个专职人员,占全馆总人数的十分之一。大量的人员从事数据库建设,势必需要对人员进行调整,为不影响图书馆传统服务,人员紧缺时需要学校通过扩充图书馆人员的编制来解决。

数据库建库的目的是追求最大限度的利用,从而发挥其社会效益和经济效益,不可重建轻用,忽视其价值体现。“一味闭门造车,求奇求怪,自娱自乐者”^[7]更应摒弃。不搞花架子,不搞形象工程,注重实用性,提倡实用与效益优先的原则。

3.2 规范管理模式,避免模糊管理

为使新闻媒体信息数据库建设管理规范化,应将数据库建设纳入学校建设规划中,以学校制度化管理来约束,学校作为唯一法人,图书馆作为学校授权人和具体实施单位,与用户单位进行合作。具体模式为:数据库建设以横向课题方式立项,将建库纳入科研范畴,将媒体数据库建设定位为横向科研课题研究。为规范经费管理,各用户单位将协议经费纳入学校科研经费中,图书馆按照科研经费管理办法对之进行支配。在广州大学《媒体广州》数据库建设实践中,这种模式得到各方认可,具有很强的规范性、科学性和操作性。

3.3 处理好知识产权问题,避免引起法律纠纷

新闻媒体信息具有时事性,按照《信息网络传播权保护条例》对时事性网络信息传播权的规定,其时事性信息内容不受版权制约,其利用具有合法性。但是,新闻媒体信息并非全为时事性信息,而且时事性信息范畴也没有法律角度的定义,因此,图书馆在建库采集这些信息时,可能会引起违反知识产权的争议。为避免纠纷,在具体信息采集过程中一定要注明信息来源,在数据网上发布时要加强权限管理,避免无保障的下载与传播,与使用单位签订合同中增加知识产权保护协议。在一些媒体信息已声明不得使用的情下,要取得使用授权才可采集利用。当然新闻媒体信息知识产权保护与图书馆使用豁免权问题争论已久,图书馆合理使用权范围到底有多大,豁免权范围如何界定,都有待大家进一步地研究与探讨,以期这些模糊空间能在《信息网络传播权保护条例》和《知识产权法》中更进一步明确。

【参考文献】

- [1]郭万召.媒体信息全文数据库建设及其应用[J].广州大学学报:社会科学版,2005(10):82-84.
- [2]Factiva@a Dow Jones & Reuters Company[EB/OL]. [2006-10-01]. <http://www.factiva.com/>.
- [3]新闻媒体数据库建设的观察思考[EB/OL]. [2006-10-12]. http://www.cnmad.com/showDetail.asp?column_id=3006.
- [4]喻国明.中国报业的困境与机遇[EB/OL]. [2005-12-08]. <http://www.qianlong.com>.
- [5]清华同方公司[EB/OL]. [2006-10-08]. <http://tpi.cnki.net/>.
- [6]于晓燕,杨宁莉.TPI 系统在校学位论文数字化管理中的应用[J].图书情报工作,2005(10):111-113.
- [7]游春山.“特色数据库”建设存在的误区及反思[J].图书馆建设,2005(2):37-38.

【作者简介】艾新革(1975—),男,中山大学资讯管理系在职研究生,副研究馆员,发表论文多篇,参编著作 1 部,广州大学图书馆项目策划推广部主任。