● 薛春香 （南京理工大学 信息管理系，南京 210094）

中文报纸文献内容深加工研究初探

［关 键 词］ 报纸文献；内容加工；文献数据库；报纸著录；报纸标引

［摘 要］ 报纸文献是一种未被充分开发的重要信息源。我国报纸文献数据库建设已经实现从题录库向全文库的发展，为报纸文献内容加工和挖掘提供了保障。但目前报纸文献缺乏统一完善的加工规范和标准，内容加工的方式也以简单的分类索引和人工剪报为主，加工自动化水平和加工深度不够，应向深层次、自动化、产品化方向发展。

［中图分类号］ G254.37；G255.3
［文献标志码］ A
［文章编号］ 1005－8214（2012）01－0014－04

报纸文献是刊登在报纸上的新闻报道、广告等一切文字和图像资料，是一种极为丰富而未被充分开发的重要信息源，具有特殊的参考价值和史料价值，被称为“活档案”。[1] 报纸文献的价值一直为专家学者所认同，但由于其加工远滞后于图书、期刊、学位论文等其他文献资料，导致其不便于查找获取，被引用率极低，其参考价值尚未得到充分发挥。据 CSSCI 统计，报纸文献被引用率一直徘徊在 3%左右，远远低于图书期刊；其中五年内报纸文献量占被引报纸文献总量的比例（即普赖斯指数）超过 60%。[2]

1 报纸文献数据库建设现状

从目前现状来看，报纸文献加工还处于大规模数字化阶段，主要体现在报纸全文浏览、版面还原技术研究，对内容的深加工研究比较薄弱，仅限于简单的分类索引和剪报应用。报纸文献数据库建设主要经历了三个阶段：（1）从纸质报纸索引向报纸题录库转变，以提供报纸文献线索为主；（2）大规模数字化加工阶段，主要是一些主流报纸的全文数据库建设和多种报纸文献混合的大型报纸全文数据库建设，以回溯建库为主；（3）数字出版阶段，各大报媒除发行纸质报纸外，还同步提供网络版、手机版报纸的浏览服务，并出现了综合性电子报纸平台，如 8 点报、AB 报、爱读爱看等等，但这些报纸平台主要提供报纸阅读功能，对内容方面的建设很少。

早期的报纸文献数据库以题录库为主，主要是从印刷版的索引文献向题录数据库发展，内容检索以提供分类索引为主，以上海图书馆的《全国报刊索引》数据库和人大书报资料中心的《中文报刊资料索引》数据库为代表。这两种索引数据库作为主要的文献检索工具，对报纸文献进行了规范的主题标引和学科分类，为用户提供了检索的便利，但不提供原文获取。

全文数据库的建设是对报纸文献进行内容深加工的前提和基础。随着全文索取需求的增长，数字化加工的规模化、数字出版和报纸网络发行，各大报纸出版集团开始回溯和建设本报的全文数据库资源，如《人民日报》图文数据库、《解放军报》图文数据库、《中国青年报》图文数据库、《经济日报》全文数据库等等。自此，全文数据库建设取得了实质性进展。

此后，一些专业文献数据库服务商开始关注报纸文献全文数据库的建设，既有综合性的中国知网《中国重要报纸全文数据库》、方正阿帕比《中国报纸资源全文数据库》等，也有专题性的如维普《中国科技经济新闻数据库》、深圳巨灵《中国财经报刊数据库》等。

从各数据库的规模来看，相较于国内目前正式在版发行的 2000 种左右的报纸种数，[3] 报纸文献数据库的加工规模远远不够；从各数据库的文献加工情况看，报纸文献的加工还处于浅层次阶段，主要提供：（1）基于报名、版名、新闻标题、作者、栏目等外部特征的检索；（2）基于全文索引的粗粒度全文检索；（3）基于简单分类索引体系的粗分类检索。个别数据库提供了基于关键词的主题检索和基于《中图法》或《中文新闻信息分类与代码》的检索与导航。

［基金项目］ 本文系教育部人文社会科学研究青年基金项目“电子报纸内容深加工研究”（项目编号：09YJC870014）；江苏省社会科学基金项目“数字报纸的自动标引研究”（项目编号：09TQC011）的研究成果之一

2 报纸文献内容深加工的主要方法

2.1 报纸文献内容加工传统方法

⑴ 索引。各种索引是传统环境下实现报纸文献检索最主要的途径，也是开发利用报纸文献的重要手段。据调查，国内正式出版的2000种中文报纸中，目前仅人民日报、光明日报、解放军报、中国青年报、解放日报、文汇报、山西日报等配置书本式索引或数据库，其他绝大多数报纸都缺乏完善的检索系统，这与网络信息时代的要求及我国新闻事业发展的现状不相适应。[4] 索引的类型以篇名索引和分类索引为主，其中分类索引主要依据《中图法》《资料法》或自编分类体系来编制，以《全国报刊索引》为代表。

⑵ 剪报。剪报的实质是将各种报纸上的信息按照专题进行采集、归类、汇总，形成全文型的资料性信息产品。剪报是信息机构针对报纸资源进行开发利用的主要方式，往往会依据本机构的服务特色就某些特定专题进行剪报。但无论是早期的手工剪报，还是现在的电子剪报，对人工的依赖程度都很大。

2.2 报纸文献的自动标引和自动分类

虽然报纸文献全文数据库建设规模越来越大，但基于全文的检索效率是低下的。因此，分类和主题标引依然是目前报纸文献内容深加工的主要形式。鉴于报纸文献信息量巨大，早在上世纪90年代就有学者提出了报纸文献的自动标引和自动分类。[5]

目前报纸文献的自动标引系统设计主要采用基于多种词表和标引源权重方案的关键词抽取标引，是一种自由标引，适当利用后控制词表进行主题规范；归类主要基于词表兼容互换原理，实现以词（串）定类；各种命名实体的抽取也是以名称词典为基础，辅以规则。由此可见，报纸文献的内容加工对各种词典、词表、类表等组成的知识组织系统依赖性很大。[6]

3 报纸文献内容深加工的主要技术问题

3.1 缺乏针对性、具体化的报纸文献加工规范

⑴ 报纸和报纸文献著录规则。报纸是一种连续性出版物，每篇报纸文献又是一个独立的著录标引对象。虽然我国有专门针对报纸期刊这类连续出版物的著录标准——《连续性资源著录规则》，但在国家标准和相关论著中对于报纸的著录标引论及甚少，大多以期刊为例进行解释说明。实际上，报纸与期刊很不相同，不能混为一谈。比如，同一种报纸存在不同地区版本、不同语种版本、不同时间版本。因此，应该有针对报纸文献加工的专门标准和规范。[7] [8] [9]

1988年，IFLA发布了一份《国际报纸编目指南》，用于规范报纸编目，但这只是一份指导性文件，并未形成报纸编目的具体规则和MARC编码标准。[10] 国内陈源蒸、石鸿飞等学者也对报纸著录中的问题进行过探讨，基本解决了报纸整体著录的问题。但时至今日图书馆和文献数据库服务商对于报纸文献的著录仍未达成共识，报纸文献数据库著录字段的设计和检索点的提供各不相同。

⑵ 报纸文献标引规范。报纸文献的标引是其内容深加工的主要形式，尤其是报纸的深度标引更是挖掘报纸文献价值、形成信息产品的主要手段。但目前缺乏针对报纸文献的标引方案和标引规则，所依据的还是通用的、简单粗略的文献主题标引规范。[11] 因此，无论是分类标引还是主题标引，受控标引还是自由标引，手工标引还是自动标引都应从便于检索、充分发挥报纸文献价值着手，针对各种性质、各种专业领域的报纸文献制订具体的标引规则和标引方案。如不同实体对象（人物、地区、机构、会议等）、不同主题（政治文献、社会新闻、经济文献、文化事业和文化活动、文艺作品、体育新闻、科技文献）、不同体裁（新闻消息、报告）、不同信息类型（广告、图片）等等，都应规定出必须标引的内容和不必标引的内容，规定出标引深度和标引专指度等，这样才能保证报纸文献价值的最大化开发和利用。

3.2 缺乏统一公认、更新及时的报纸文献知识组织工具

各种分类表、主题词表、术语表等知识组织工具在文献内容加工组织和开发利用中具有重要的支撑作用。但目前，针对报纸文献的各种词表存在编制困难、更新滞后、难以统一普及、缺乏互操作性等一系列问题。[8]

⑴ 分类表。在《中文新闻信息分类与代码》标准发布之前，报纸文献的分类体系一直是各自为政，比如知网《中国重要报纸全文数据库》先是采用自编的三级类目体系，包括10大专辑、168个专题、近3600个细目，后又改用《中图法》类号标注；而《全国报刊索引》数据库则以《资料法》作为分类依据；各大报系又有适应本报内容的自编分类体系，缺乏针对新闻信息特点的专用统一的分类体系。2003年科技部启动《中文新闻信息技术标准》的国家科技攻关项目，形成了新闻信息分类标准——《中文新闻信息分类与代码》，并于2006年5月付诸实施。该标准把政治、经济、文化三大部类作为一级类目划分的基础，采用层次编码法，主表从粗到细，划分出23个一级类目、

315 个二级类目、5683 个细目，类目总数达到 9314 个、类目层级达到 5 级，同时附加了 6 个通用复分表和 11 个专类复分表。《中文新闻信息分类与代码》国家标准的颁行推动了报纸文献分类组织的统一，但限于人力、分类体系转换成本和效率等诸多原因，普及度和采用率并不高。

⑵ 主题词表。报纸文献涉及的主题、体裁甚广，并且不断有新主题、新事物涌现，很难用一部通用的主题词表来覆盖。《全国报刊索引》以综合型《中国分类主题词表》作为其主题标引的受控依据；新华社则专门编制了用于存储和检索新闻资料的专业叙词表——《新闻叙词表》，收录正式主题词 8603 条，非正式主题词 1201 条，学科范围涉及国内外政治、军事、外交、文化、科技及社会生活各个方面。但总体来说，由于报纸文献主题标引规模较小，即使标引也以自由标引为主，因此，适用于报纸文献的主题词表编制和应用研究甚少。

除了分类表和主题词表外，因为报纸文献中有大量的新闻报道，其中的人名、国家地区、事件名、机构名、产品名等等命名实体都具有一定的检索意义和参考价值。为了对这些命名实体进行抽取和规范控制，还需要名称权威档等知识组织系统的支撑。

3.3 缺乏特色性、高增值的报纸文献内容深加工方式

从目前报纸文献内容加工的方式来看，仍以传统文献著录和标引，形成指示性文献检索线索为主，或是人工依赖程度很大的剪报产品，内容深加工形式单一。

报纸文献涉及范围广泛，既有新闻报道类的消息型信息，也有资料型信息，还有知识型信息；报纸文献的受众面广，用户特点和用户需求各异。因此，完全可在及时、新颖且多为第一手资料的报纸文献基础上形成针对性、特色性、高增值的各种信息产品。

⑴ 专题库。按照各种实体、具体事件、具体行业、具体领域整合多种报纸上一定时间段内的各种相关文献，形成各种专题数据库，提供给不同用户。

⑵ 知识库。从抽取各种事实性、数据性的报纸资料中抽取事实、数据、实例等形成知识库，即事实数据库产品。

⑶ 参考咨询库。专题库和知识库还只是基于报纸文献一手资料的采集、选择和撷取的加工，而在这些分类别、序化的聚合信息基础上，辅以数据挖掘和专家智慧，则可以形成研究性、预测性的市场调查报告、行情分析、趋势预测等高增值的信息产品，使公开的报纸文献成为重要的竞争情报信息源。

4 报纸文献内容深加工的主要趋向

无论是旧报纸还是现行报纸，单纯的数字化是远远不够的，必须实现报纸文献内容的深加工，形成增值信息产品。目前学界、业界对于网络新闻的组织、挖掘探索越来越多，虽然网络新闻并不等同于报纸文献，但将在网络信息挖掘、图书期刊论文资料等领域内容加工的方法和技术移植到报纸文献内容加工领域，并结合报纸文献的特点形成针对报纸文献内容加工的专门方法是值得尝试的。具体如下：

⑴ 由各自为政的分类索引向基于新闻分类标准整合报纸信息资源方向发展。分类索引是报纸文献内容组织最主要的传统方式，但由于缺乏统一的分类体系，导致各个报系和文献数据库之间分类组织互操作的障碍，更遑论进行资源整合。现在作为国家标准的《中文新闻信息分类与代码》分类表已经颁行，但让各单位立即摒弃原有的分类体系却不可行，何况这个国家标准的适用性还需要进一步的验证。因此，从资源整合的角度出发，可考虑在沿用原有分类体系的基础上，将其与标准分类表之间进行映射转换，通过分类表的互操作来实现资源整合。

⑵ 由简单主题标引向各种实体抽取方向发展。实体标引在报纸文献标引中是有历史的，而各种命名实体的抽取和标注对于报纸文献检索、建立文献关联、形成专题产品都具有重要意义。因此，在计算语言学和信息组织智能化不断发展的前提下，报纸文献的主题标引还需强化，并且要进行多元、多角度、全方位的深度标引。

⑶ 由传统剪报向个性化、专题化信息产品方向发展。剪报是在报纸文献基础上形成的一种增值性信息产品，传统的“剪刀加浆糊”的工作方式已经不能适应快速精准的现代信息需求。在报纸文献有序组织、深度揭示的基础上，对用户信息需求进行细化，实现报纸文献信息的重组和创新，从而形成个性化、专题化的剪报产品。

⑷ 由传统文献组织向内容挖掘方向发展。报纸文献的内容加工不能局限在为提供检索服务的信息序化层面，而应向内容挖掘层面进行深加工。报纸文献的内容挖掘既包括单篇文献中的主题揭示、各种命名实体的抽取和语义关联、观点挖掘等；还包括集合文献的专题聚类、热点追踪、观点导向分析、新闻过滤、舆情预警等等，真正发挥报纸文献的喉舌、参谋作用。

● 陈钦明 (泉州师范学院 图书馆，福建 泉州 362000)

浅析我国数字图书馆“御敌”策略

——Google 数字图书馆“兵临城下”

[关键词] 数字图书馆；Google；版权；利益平衡；融资

[摘 要] 通过分析 google 数字图书馆发展现状及其战略目标，结合我国数字图书馆发展存在的问题，探索我国数字图书馆走出目前困境的方法，提出解决作者、出版商、读者利益动态平衡问题的方案及一系列提高我国数字图书馆竞争力的组合措施，在与 google 数字图书馆竞争中，发挥我国的市场规模优势，走出国门，利用我国的文化潜力，进军国际数字图书馆市场，提升国家软实力，提高国家竞争力。

[中图分类号] G250.76
[文献标志码] A
[文章编号] 1005－8214(2012)01－0017－04

1 我国数字图书馆及 Google 数字图书馆发展现状

1.1 Google 数字图书馆现状

Google 数字图书馆计划想依靠其先进的技术和雄厚的资金支持，计划用 5 年时间扫描 5000 万册图书并上网免费供读者阅览，此计划得到了广大读者的支持。斯坦福大学图书馆馆长 Michael A.Keller 对 Google 数字图书馆评论说：“多年来，图书馆一直在努力数字化图书，但限于技术和资金的双重原因，速度是非

* 本文系泉州师范学院校自选课题“从 Google 数字图书馆发展现状探索我国数字图书馆发展策略” (项目编号：2010SK30)

[参考文献]
[1] 张琪玉. 报纸文献是一种极为丰富而未被充分开发的信息源—关于发展报纸文献索引和数据库的思考 [J]. 图书馆杂志, 1999 (2)：7－8.
[2] 王智琦，李秋实. 基于 CSSCI 不同类型文献的发展趋势定量研究 [J]. 图书馆, 2008 (3)：38－40，68.
[3] 中华人民共和国新闻出版总署. 2009 年全国新闻出版业基本情况 [EB/OL]. (2010－09－07) [2011－06－11]. http://www.gapp.gov.cn/cms/html/21/493/201009/702538.html.
[4] 葛永庆. 开发报纸文献的重要手段和有效途径—兼谈《申报索引》的编纂出版 [J]. 中国索引，2008 (2)：2－3.
[5] 宋明亮. 报纸文献机助自由标引研究及对汉语后控制词表动态维护的思考——《解放军报》模拟检索系统设计实验报告 [D]. 中国人民解放军空军政治学院硕士论文，1994.
[6] 辛乘胜. 人民日报新闻文献自动标引系统的设计与实现 [J]. 中国传媒科技，1997 (3)：17-19
[7] 李素建. 人民日报标引系统 [EB/OL]. (2002－11－18) [2011－06－12]. http://www.icl.pku.edu.cn/member/lisujian/papers/人民日报标引系统 intro.pdf.
[8] 查贵庭，侯汉清. 基于多词表的自动标引技术研究——新华社新闻稿自动标引的实验 [J]. 情报学报，2002，21 (3)：273－277.
[9] 马金林.《申报》全文数据库的自动标引 [J]. 信息系统工程，2009 (11)：39－40.
[10] Hana Komorous，Robert Harriman. International Guidelines for the Cataloguing of Newspapers [EB/OL]. (1988－07－01) [2011－06－11]. http://www.ifla.org/VII/s39/broch/intguide.pdf.
[11] 许斌. 关于开发报纸文献索引及数据库的思考 [J]. 图书馆学研究，2005 (2)：41－42.

[作者简介] 薛春香 (1979－)，女，南京理工大学信息管理系副教授，研究方向为：智能信息组织、知识组织系统构建。

[收稿日期] 2011－08－01 [责任编辑] 宋玉军