文章编号：1005-8451（2007）10-0038-03

铁路科技期刊全文数据库系统的研究和建设

蔺红生，延秉真，李 梦
（铁道科学研究院 信息技术研究所，北京 100081）

摘 要：简要介绍全文检索技术的发展状况、最新进展，针对铁路科技期刊全文数据库的需求，详细讨论全文检索系统的体系结构，系统功能及建设流程，该系统的应用必将大大提高铁路科研人员获取原文的及时性和方便性。

关键词：全文检索；铁路；科技期刊；数据库
中图分类号：TP39 文献标识码：A

Research and construction on Railway Science and Technology Journa Full Text Database System

LIN Hong-shen, YAN Bing-zhen, LI Meng
(Scientific and Technological Information Research Institute, China Academy of Railway Scierces, Beijng 100081, China)

Abstract: It was briefly introduced the development, the newest progress of the full text retrieval technology, and according to the requirement of the railway science and technology journal full text database, discussed in detail the full Text Retrieval System architecture, the function and the construction process. The application of the System would greatly improve the convenience for the railway researcher to gain the original text.

Key words: full text retrieval; railway; science and technical journal; database

随着信息化进程的深入开展，各行业都建设了大量的数据库，已经基本解决了数据信息的有序化和存储管理问题，但目前信息资源的共享和利用水平还比需要进一步提高。从技术角度来看，也大都是建立在关系型数据库基础之上，传统关系型数据库能对结构化数据提供简便的管理和查询手段，但无法有效处理大量的非结构化信息，如科技文献、图书目录、Web页面、新闻资料、档案文件、专利、法律、项目文档、合同和技术文档等，这类信息已占有整个信息量的80%以上。对于结构化数据，用RDBMS（关系数据库管理系统）技术来管理是目前最好的一种方式。但是由于RDBMS自身底层结构的缘故使得它管理大量非结构化数据显得有些先天不足，特别是查询这些海量非结构化数据的速度较慢。而通过全文检索技术就能高效地管理这些非结构化数据。

1 全文检索技术及其发展状况

1.1 全文检索技术

收稿日期：2007-04-18

作者简介：蔺红生，研究实习员；延秉真，副研究员。

全文检索（Full-Text Retrieval）：以各类非结构化数据诸如文字、声音、图像等为处理对象，包括信息的存储、组织、表现、查询和存取等各个方面，提供按照数据资料的内容而不是外在特征来实现的信息检索手段，其核心为文本信息的索引和检索，常用的实现全文检索的方法主要有以下两种：

第1种方法是不对数据库建立索引而直接对文章进行匹配的方法。这种方法没有建立索引库，因此，所占空间较少，但同时正是因为它没有索引库，所以在进行全文匹配时要花费大量的时间。

第2种方法则是一种为全文建立倒排索引库的方法。这种方法可以大大节省检索的时间。但同时，这种方法需要占用一定的存储空间来建立索引库。

目前，绝大多数全文检索系统，均采用“空间换时间”的策略实现全文检索，即本文中提到的通过建立索引的方式实现全文检索。

1.2 评价全文检索系统的指标因素

（1）查全率，即系统在进行某一检索时，检索出的相关资料量与系统资料库中相关资料总量的比率；（2）查准率，是保证我们找到最有用资料的一个关键，是系统在进行某一检索时，检索出的有用资料数量与检索出资料总量的比率；（3）检索速度

或者说响应时间，指的是从提交检索条件到查出资料结果所需的时间。检索速度的提高取决于检索方式、索引技术、数据库技术和编程技术等多方面。此外还有诸如检索并发用户数、输出形式（输出信息表现形式）、系统的可靠性和稳定性等指标也是衡量全文检索系统优劣的指标要素。

1.3 全文检索技术最新进展

目前，国内外对全文检索的研究可以说是达到一个高潮。许多研究机构和商业组织都在进行这方面的研究。

在众多的中文全文检索系统中，其最新的进展表现在：

（1）采用先进的中文信息处理技术，内嵌汉语自动分词系统，支持按词索引、按字索引、按关键词索引、字词混合索引，大大提高了检索的准确性和响应时间。

（2）检索信息快、准而且基于优化的查询算法，使得G级数据库查询速度达到亚秒级，并支持大量并发用户同时访问。允许使用文中的任意字、词、句和片段进行检索，提供了基于文献内容而不仅仅是文献外部特征的全文检索手段。TRS所提供的按词和按用户自定义关键词进行索引和检索，以及基于知识词典的扩展检索功能，满足了特殊应用领域的高查准率和高查全率的要求。

（3）检索服务更趋智能化、个性化。从用户分类、行为、内容等方面的不同需求，为用户进行定制、细分检索服务。通过记录完整的访问日志，并对其进行细致的分析，可以帮助用户更加准确地获得客户访问信息，通过对客户的访问痕迹进行统计分析，对应用中相关内容进行调整。

（4）提供分布式检索和支持负载均衡功能，满足大数据量和高并发的检索要求。

2 铁路科技期刊全文数据库系统

2.1 系统总体体系结构

为了保证整个系统具有良好的易维护性与易扩展性，全文检索系统采用了B/S三层架构，自下而上，可以分为3部分—数据层、业务逻辑层、表现层如图1。三层架构方式不但保证了系统具有良好的可维护性、可扩展性，而且能够有效地保证系统数据的安全性。全文检索系统采用面向对象的J2EE技术进行设计和实现，完全跨平台，可以在各种主流硬件平台和操作系统上运行。

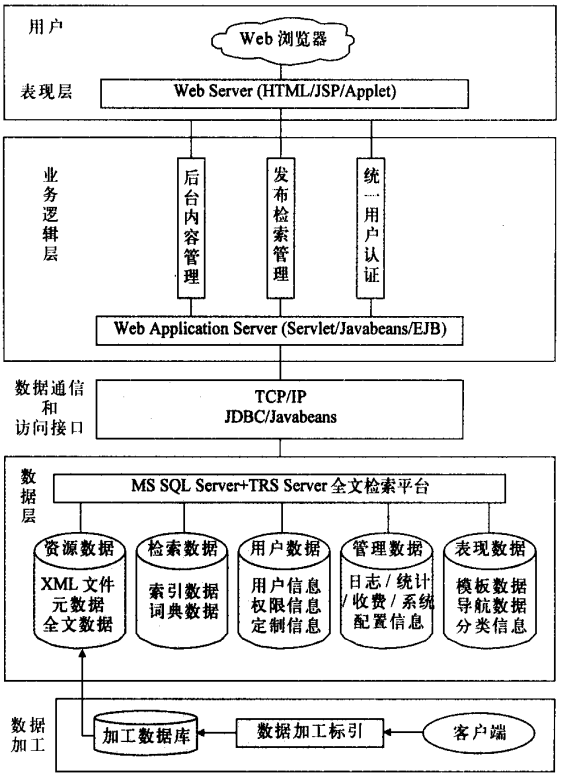

图1　全文检索系统总体体系结构图

2.2 系统功能结构

从功能模块角度描述铁路科技期刊数据库系统，可以分为：数据资源建设模块、数据资源管理模块和信息发布服务模块。如图2所示。

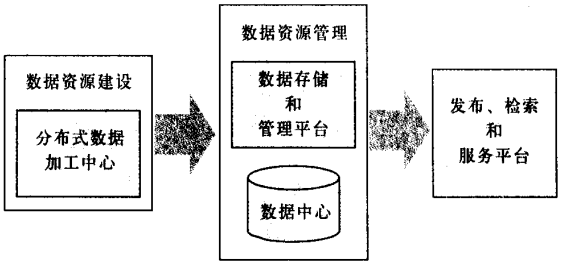

图2　全文检索系统功能结构图

信息资源的加工和数字化工作是全文数据库建设的基础，该模块通过将大量铁路科技期刊等文献资料进行采集、加工、整理、处理和分析，逐渐形成强大的后台资源中心。

数据资源管理模块采用全文数据库技术作为后台强大的内容处理引擎，把铁路科技期刊专题资源

数据库作为强大的后台资源中心，实现对铁路科技期刊的及时发布、精确控制、调整更新等的管理功能，为用户最终提供丰富的、有价值的期刊全文信息。此外，还包括索引管理，用户管理，日志管理等功能，

信息发布模块是系统功能中重要组成部分，本模块对信息进行分类，系统化、标准化发布到门户网站上，是一个高质量、高效率、智能化的信息门户，是用户浏览信息的重要工具。信息发布模块是基于浏览器 Web 操作模式的应用服务，管理控制台完全基于 Web 方式，实现了远程管理，使管理与维护更加方便灵活。

2.3 铁路科技期刊数字化加工流程

纸质期刊通过扫描、OCR 识别、标引和入库等环节，实现电子期刊的全文检索，其加工流程如图 3。

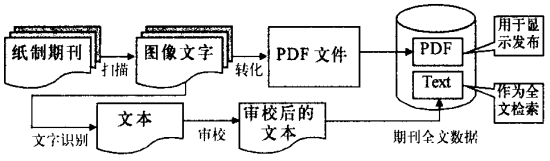

图 3 铁路科技期刊数字化加工流程示意图

（1）扫描：将需要加工的期刊在扫描仪上进行扫描，得到的是图像文件，以 TIFF、JPG 等图像格式保存。

（2）图像预处理：对扫描后的图像进行预处理，包括倾斜校正、去除噪声点、图像的版面分析等。

（3）文字识别：通过 OCR 软件进行文字识别，目前 OCR 软件的识别率在 90%以上。

（4）校对：通过人工校对方法，确保最终文本的差错率在一定范围之内。

（5）版面恢复：将识别校对后的文本导出生成双层 PDF 格式的文档，双层 PDF 即 PDF 文件的每一页都包含两层，上层是从纸质文件扫描出来的经过图像预处理后的图像，下层是用 OCR 软件对扫描图像进行识别后产生的文字结果。这样用户在阅读 PDF 文件时看到的是扫描图像，可以 100% 保留原始版面效果，在需要的时候，又可以通过下层的文字信息支持选择、复制、检索等功能。

（6）标引信息形成和入库：对双层 PDF 文档进行文档部分信息的自动标引，比如：文章的“标题”、“作者”、“关键词”、“摘要”等信息，部分标引信息形成则需要通过人工的干预，标引完成后，即可批处理汇入全文数据库。

（7）质量控制：质量控制是数据高效、高质录入的必不可少的环节，是检验最终文本错误率的有效工具，贯穿期刊数字化处理的整个过程。

2.4 系统应用情况

铁路科技期刊全文数据库系统的建设取得了良好的应用效果，该系统的应用必将大大提高铁路科研人员获取原文的及时性和方便性。

系统提供了丰富的内容检索方式，提供多种检索入口。针对不同的内容来源，可以灵活的使用，确保用户可以准确迅速地找到相关的内容。全文检索系统采用内嵌汉语自动分词系统，支持按词索引、按字索引、按关键词索引，并统计建立了大量歧义排除规则，有效提高了分词准确性，在输入检索词对文档进行全文检索的时候，首先对索引库进行检索，然后定位到原文档，极大地提高了文档的查询速度和查全率。用户可以按照数据库的字段进行精确的组合检索。如按标题、作者、刊名和关键词等字段进行组合检索，大大提高了文档的查准率。此外，系统还提供智能检索，是对关键词检索的扩展功能，例如输入“铁路”检索，检索结果中可以命中包含“铁路”和“铁道”等的记录。

3 结束语

“十一五”期间，我国铁路进入快速发展阶段，广大铁路科研工作者在科研过程中需要更加及时、深入的科技文献资料作为参考，紧紧抓住我国铁路建设发展的黄金机遇期，充分利用全文检索及网络数据库技术，对铁路行业的科学知识、技术、管理经验进行系统化、理论化加工、研究、传播和普及，通过铁路科技期刊全文数据库系统平台，为铁路行业科技创新提供高效、权威、准确的信息服务。

参考文献：

[1] 姬 红. TRS 全中文检索系统的应用[J]. 现代电子技术，2003 (21).

[2] 牟有静，侯丽梅. 浅谈数字图书馆与全文检索技术[J]. 情报科学，2002 (5).

[3] 王志刚，唐文忠，杨宗煦. 用 JSP 和 TRS 开发文献管理系统[J]. 计算机应用，2001 (8).

[4] 黄 长. 利用 TRS 全文检索系统建设专题数据库的研究和实践 [J]. 图书馆论坛，2005 (3).