鑫磐全息数据库和现代图书馆文献信息服务

王怀汀 毛 军

（北京鑫磐软件技术有限公司 北京 100029）（中国科学院文献情报中心 北京 100080）

【摘要】 为了在Internet和Intranet上提供数据库网络查询服务，图书馆需要将现有的数据库进行加工（如转换成全文数据库）。鑫磐公司的鑫磐全息数据库GSADB，为文献情报机构提供网络数据库查询服务提供全面解决方案。本文论述了GSADB的数据库定义、索引、因特网互连等关键技术和应用。

【关键词】 数据库 文献信息服务

Goldsoft all Information Database (GSADB) and Modern Library Information Service

Wang Huaiting

(Beijing Goldsoft Corporation, Limited, Beijing)

Mao Jun

(The Document and Information Center of Chinese Academy of Science, Beijing)

【Abstract】 Online database retrieval service gradually attracts modern library's attention. More and more libraries offer online information service including full—text information retrieval. With Goldsoft(tm) ADB, Libraries can not only easily connect their databases into Internet/Intranet, but also quickly build their own full—text database. This article describes some technological features of Goldsoft(tm) ADB and presents our database solution to library information service.

1 前 言

数据库技术自50年代兴起，经历了三个发展阶段：首先是50年代的文档型数据库；其次是60、70年代的网状数据库和等级数据库；最后到80年代则是关系数据库的天下。目前商品数据库市场上关系数据库占有很大的份额（约80%）。90年代以来，面向对象的数据库越来越多地出现在各种数据库产品中（如ORACLE）。相关的国际标准如SQL 3业已成熟。图书馆利用数据库进行文献信息服务也经历了一个漫长的过程，如在图书馆业务工作中对文献采购、编目、流通等过程进行自动化管理，随着因特网在国内的普及，图书馆需要提供传统的书目数据服务以外的全文数据库查询服务以及数据库的因特网发布，因此利用新的数据库技术，就成为一个紧迫的任务。

收稿日期：1998—12—25

2 现代图书馆文献信息服务中的应用

2.1 专题数据库和全文数据库

图书馆提供的传统数据库查询服务大多限于目次的级别，读者仍需查阅印刷型文献，电子文献的出现使读者要求不但能够通过图书馆自动化系统查阅文献的目次信息，还希望直接查看其文摘、前言、甚至全文；对于专业图书馆，经常的需求表现在读者需要了解某个专题的所有文献和最新文献，包括目次、文摘、全文甚至多媒体资料。这些都需要图书馆提供专题数据库和全文数据库。

2.2 数据库的网上发布和查询

因特网在国内的普及，读者可以在任何连通因特网的地方查询图书馆的各种文献数据，而不是象从前那样必须到图书馆的查询工作站去查询图书馆的馆藏，如何快速地将图书馆的文献数据库联入因

特网供读者查询，这同样需要在原有的关系数据库基础上使用新的数据库技术。

2.3 多媒体数据库和海量数据

数据库的内容现在不再局限在文本数据，而应该包括视频、音频、图片、表格等多媒体数据，这是传统的关系型数据库难于处理的；而且数据库的容量也有很大的扩充，记录数达到上千万条，甚至上亿条，如化学文摘(CA)等光盘数据库。提供多媒体文献信息服务将是图书馆的一项新的服务内容。

3 传统的关系数据库在新的文献服务方面的局限

关系数据库在 80 年代已经趋于成熟，图书馆在处理文献领域也有相关的国际标准如 LC－MARC、UNI－MARC、CCF 以及 ISO 2709 但是这些标准和协议大多针对的是印刷型文献，并且只限于二次文献(如图书、期刊等)，随着因特网的普及和电子文献的增加，图书馆的文献的级别将延伸到一次文献，甚至全文和多媒体。这时关系数据库在处理非结构化文献数据和海量数据方面就存在如下不足：

3.1 定长数据

关系数据库的数据类型实际上是一个二维表。每一行由若干数据项(列)，数据项的长度和类型是预先定义好的，这就使该数据项的灵活性较差，如果实际数据的长度小于定义的长度，则浪费了空间；反之则会出现数据的丢失，虽然关系数据库采用一定的手段来处理变长的数据项(如 MEMO 类型)但是在抽取检索项建立索引又会出现问题。例如：图书馆中如果希望处理文摘数据(属变长数据)并且对其进行检索输出，就难以通过关系数据库来完成。

3.2 索引技术

关系数据库对定义好的定长数据项可以建立索引进行查询，但是却不能直接对数据项中的重复子字段和重复字段建立索引，而是通过建立一对多和多对多数据库加以解决。这种一对多和多对多关系数据库在进行关系联接运算时效率不高，在库记录超过一定数量时，数据库查询会变得明显缓慢，严重影响系统的检索效率。

3.3 全文检索

从上述的索引和数据结构可以看出，关系数据库难以处理变长数据，其索引方式有限，因此对文献进行全文检索是很难的。特别是对汉语文献进行全文检索，还面临着一个汉字的切分问题，很容易产生庞大的索引文件。在采用 C/S 结构的数据库查询系统中，关系数据库处理全文检索几乎不可能。

4 鑫磐全息数据库 GSADB 的优势

1998 年 12 月，鑫磐软件技术有限公司在北京发布了新一代全息数据库管理信息系统：鑫磐全息数据库 GSADB (GoldSoft All information DataBase)和因特网全息数据库系统 GSADB WEB，标志着鑫磐公司在 Internet 和 Intranet 上处理非结构化信息、全文信息、多媒体信息和其它海量信息领域占据了领先位置。GSADB 较好地解决了上述关系数据库的缺陷，能够为图书馆开展进一步的文献信息服务提供强有力的技术支持。其主要特征如下：

4.1 各种类型数据库的快速接入

GSADB 首先保证对各种关系型数据库的良好兼容能力，从小型数据库管理系统如：Dbase 系列，FoxPro，Paradox，Access 等；到大型的数据库如 Oracle、Sybase 和 Informix 等，通过 GSADB 的数据库引擎都能迅速挂接到系统中，以便进行深入的加工和处理。

4.2 多种索引和灵活的检索技术

GSADB 采用 B＊结构进行倒排档文件的维护，大大提高了数据库的检索效率，确定一个检索词的位置最多需要 7 次 I/O。系统采用八种索引方式，支持对整个字段、子字段甚至全文检索。其检索技术丰富多样。既有传统的精确词检索、前方一致检索、相关词检索；也有最新的定题检索、回溯检索和全文检索等。如 GSADB 支持字典检索，用户可以直接从检索词列表中选择检索词，形成检索式，完成检索功能；检索词典可以按照用户要求进行分类，形成分类字典，方便检索。

4.3 全文检索

GSADB 的突出特色就表现在其全文检索的功能上，特别是在因特网的环境下，GSADB 允许用户将数据库中具有查询意义的任意内容作为可检索词。对于西文数据因为着每一个单词和数字都是可检索的，对于中文数据，则任意组配词都是可检索的。GSADB 对中文数据进行全文索引时，采用了单汉字索引技术，查询时根据用户提出的检索词，自动

进行物理位置的组配处理。

4.4 变长数据和海量数据

GSADB 利用自定义的数据库格式处理变长数据，数据库的记录是按变长存储，记录内也是变长的，普通字段可以作为多媒体字段处理，存放声音、图象等多媒体信息。管理的数据库数量庞大；每个数据库的最大记录数可以达到 1000 万条；每条记录的最大长度可以达到 32000 个汉字；每个数据库最多可以有 800 个字段；每个字段的最大长度可以达到 32000 个汉字。记录与字段的最大长度可以根据用户的特殊要求定制。

5 鑫磐软件为图书馆提供的文献信息服务的解决方案

鑫磐软件提供的解决方案分四个部分：Internet 和 Intranet 网络环境；网络操作系统和因特网应用服务器软件；数据库和网络管理人员培训。可以帮助用户成功地实现从图书馆到因特网上图书馆的转变。

5.1 Internet 和 Intranet 的网络环境

鑫磐公司提供网络的结构和配置方案，以协助网络公司完成具体的布线和连网的调试。在局域网方面，鑫磐软件提供总线型、星型等拓扑结构的网络布线方案，服务器和工作站的硬件配置参数(CPU、内存、网卡、硬盘等)；在因特网方面，提供 WWW 服务器、电子邮件服务器的网络连接配置参数(路由等)。对于用户现有的局域网环境，鑫磐公司会提出优化的方案。

5.2 网络操作系统和因特网应用服务器软件

局域网网络操作系统建议在服务器端用 Microsoft 的 Windows NT 4.0 或 Novell 的 Netware，工作站端可以使用 Windows 95，Windows 98，Windows NT for Workstation 等操作系统；1999 年年中微软将推出 Windows NT 5.0，是时 NT 在安全性和灵活性将可以同 Unix 操作系统抗衡；在因特网服务器软件方面可以配置 WWW 服务器软件如 Microsoft Information Server，Netscape FastTrack 等；电子邮件服务器如 Microsoft Exchange Server 等；还有其它的 FTP、TELNET 服务器。

5.3 数据库

主要采取 3 种途径建立图书馆的数据库：第一是对现有的文献数据库进行改造，可以通过 GSADB 直接挂接到系统中；也可以对现有的数据库增加新的文摘和多媒体内容；第二种方式是购买商品化的数据库如美国的化学文摘 CA、清华大学的学术期刊光盘等，将其转入 GSADB 中进行处理后便可提供查询服务；第三个方法是自己建立专题文献库。根据图书馆的实际情况，鑫磐公司会提供可行方案。采用 GSADB 和 GSADB WEB，图书馆可以迅速开展 Internet 和 Intranet 文献信息查询服务。

5.4 网络管理人员培训和技术支持

图书馆利用 GSADB 和 GSADB WEB 提供 Internet 和 Intranet 文献信息服务，对文献数据库还有一个维护和更新的问题。网上的数据库内容更新、检索点增加、新的检索式的构造，都需要网络管理员的参与。鑫磐公司对 GSADB 的用户提供免费的技术培训，保证用户能够熟练的掌握 GSADB 的使用方法，另外还会通过召开系统管理员培训班、热线电话咨询、网络咨询和上门服务的方式来保证 GSADB 的顺利运行。

提供 Internet 和 Intranet 文献数据库查询服务，是现代图书馆深化自己的文献服务内容，突出自己的文献特色的好办法。在全国和地区文献资源共享网络中，图书馆可以通过这种服务方式提高自己的地位，鑫磐全息数据库在这方面为图书馆用户提供了一个功能强大、快速便捷的工具，帮助图书馆在自动化和网络化的进程中更加得心应手。

参考文献

1 沈玉兰. 新型书目数据格式研究. 情报学报 1998,(5):332－340
2 戴诗勇 吴广印. 中文浏览器:INTERNET 全球通的开发. 情报学报, 1998,(5):341－346
3 丁 蔚. 单汉字检索系统后控词表的改进研究. 现代图书情报技术,1998,(5):25－28