信息系统工程 CCNEWS
国家经济信息系统核心期刊
面向信息工程行业的专业理论媒体
国家经济信息系统核心期刊 中国核心期刊（精选）数据库入选期刊 CNKI中国期刊全文数据库收录期刊

刊物基本信息

国际标准刊号 ISSN1001-2362
国内统一刊号 CN12-1158/N
主管单位：天津市发展和改革委员会
主办单位：天津市信息中心
支持单位：国家信息中心
顾　　问：高新民　周宏仁　徐漳河
　　杜　链　胡小明
编委会主任：沈大风
编委会副主任：张晓雁　王华峰
社　　长：高广田
总　　编：王华峰
副 总 编：王颖振
执行主编：高铭铎
编辑部主任：施　洋
编 辑 部：林仲信　李海京　黄夜晓
　　王　雪　尹正富
编辑出版：《信息系统工程》杂志社
地　　址：天津市河西区友谊路39号
邮　　编：3000201
北京组稿中心
地　　址：北京市朝阳区建国路15号院甲1号华文国际传媒大厦B座732室
邮　　编：100024
联 系 人：施洋
电子信箱：xxxtgc@126.com
刊　　期：月刊
邮发代号：82-173
国外代号：M8041
国外发行：中国国际图书贸易总公司
总 发 行：北京报刊发行局
　　全国各地邮局
印　　刷：北京北方印刷厂
广告经营许可证：1201034000019

非结构化数据库及其应用分析

何淑娟

（南通农业职业技术学院信息系 江苏 南通 226007）

【摘要】介绍了非结构化数据库的概念。分析了非结构化数据库在存储机制和索引机制上的变革及作用。提出了在多类型文档管理中使用非结构化数据库的思想，并分析了相关应用中的若干关键问题。

【关键词】非结构化；数据库；多类型文档

1 非结构化数据库的基本概念

所谓非结构化数据库，是指其字段长度可变，并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据库。它从数据模型入手，采用子字段、多值字段以及变长字段的机制，允许创建许多不同类型的非结构化或任意格式的字段，从而突破了关系数据库严格的表结构，解决了关系数据库模型过于简单、不便表达复杂嵌套的问题；在其底层存储机制的变革基础上，采用先进的倒排档索引技术，从而实现了对于海量文献信息的快速全文检索的功能，并同时支持多种字段限定检索。对于多媒体信息的存储和管理，非结构化数据库系统采用外部文件方式，摈弃了传统关系型数据库采用二进制字段存储的方式，实现了对于图形、声音等多媒体信息的高效管理[1]。 其高效性具体表现在：

（1）在数据库结构定义上 非结构化数据库系统实现了对于变长字段、重复字段和子字段的定义、存储和管理，并且记录的数目、长度，字段数目与长度以及字段可重复次数均可不受限制，允许数据项具有多值性和可包含子字段，充分满足了图书馆建立文献数据库的特殊管理要求。

（2）在数据著录格式方面 非结构化数据库不仅支持国际标准（ISO-2709,MARC,CCF）和国内标准(CCFC)格式，而且支持最新的SGML和XML格式，具有可扩展性，可以与其他元数据单元(项目)连接使用。在数据处理对象上，非结构化数据库采用面向对象技术，不仅可以处理TXT文本、DOC、EXCEL、PPT、PDF、S2、PS2等流行的文件类型数据，而且可对图象、音频、视频、计算机程序以及网址资源进行编目和数字化处理，覆盖了多类型文档应用领域内几乎所有的文献数据类型。同时，非结构化数据库支持外挂文件的全文检索，其独特的外部文件支持能力使图书馆能轻松实现二次文献挂接全文的功能。

（3）在信息检索查询方面 非结构化数据库内嵌全文检索引擎，采用倒排档索引技术，不仅能够对整个字段进行查询，而且可以提供子字段、关键词、自由词、标引词、位置词和全文任意词的单项及组配检索。而且速度也非常快，一般不受文献量（上千万条记录）的影响，满足海量数据检索的需要。

（4）在信息查全率和查准率方面，非结构化数据库采用自然语言处理和人工智能技术，提供基于内容的检索和ANY词检索方式，并在检索中实现对于特定类目相关词的利用，大大提高了系统的查全率。同时非结构化数据库支持的禁用词，可以过滤掉一些没有检索意义的英文虚词如“I、 TO”等，以提高查准率。

（5）非结构化数据库提供了后控制词表检索系统。后控制词表系统是提高自然语言全文检索效率，减轻用户负担的有效途径。该系统对于后控制词表采取数据库管理方式，与全文检索的检索式构造相连接。对每一个检索词提供用（UF）、代（USE）、属（BT）、分（NT）、参（RT）等关系词，用户可根据具体检索需求选取关系词，并将之增加到检索表达式中，从而实现检索表达式的优化，提高全文检索的效率，降低用户负担 [2]。

2 非结构化数据库的应用

关系数据库从设计之初并没有也不可能考虑到以HTTP为基础、HTML为文件格式的因特网的需求，只是在因特网出现后才作出相应的调整，因此关系数据库在基于网络应用时由于结构模型等原因的限制，不能与网络完全融合，需在网络与数据库之间加入大量的中间件，从而在无形中加大了数据库基于网络应用的难度。并且对于网络环境下网络应用，如各种非结构化文档信息、多媒体信息以及全文检索需求显得力不从心。虽然后来关系数据库对于这些需求作出了一些适应性调

整，但对于网络环境下网络应用不可或缺的检索效率、全文检索能力等却无法解决[3] [4]。同时，关系数据库的基于中间件的解决方案又给网络应用带来了新的网络瓶颈，应用服务器端由于与数据库频繁交互，因其本身的效率和数据库检索的效率造成网络应用在应用服务器端的阻塞。

（**1**）文档型和多媒体数据类型的支持问题

在网络应用中，存在大量的复杂数据类型，如各种超文本文档信息，各种图片、声音等多媒体信息资源，如何对这些信息资源进行有效的存储、管理、检索，是网络数据库必须解决的问题，iBASE非结构化网络数据库系统通过其外部文件数据类型，可以管理各种文档信息、多媒体信息，并且，对于各种具有检索意义的文档信息资源，如HTML、DOC、RTF、TXT等还提供了强大的全文检索能力。

（**2**）数据库的全文检索问题

在网络应用中，如何从浩瀚的信息海洋中查找到所需的信息，如何保证所查询信息的全面性和准确性，也是一个网络数据库应用必须解决的问题。非结构化网络数据库系统通过其独特的索引技术和基于布尔检索表达式的查询检索算法，解决了基于字段级和数据库级的全文检索问题，用户可以针对数据库中特定的字段也可针对整个数据库进行全文检索，从而从数据库中检索出感兴趣的内容 [5]。

（**3**）网络数据库应用中的查询和检索效率问题

作为网络应用，由于需要面对大量的用户群和大量的瞬时并发数据库查询检索，其数据库查询和检索效率就是一个极其关键的问题。iBASE非结构化网络数据库系统主要通过重复字段和子字段来保证数据库查询和检索的效率，实现了数据库的一条记录中一维表和二维表嵌套，从而避免了关系数据库在大数据量时由于表连接查询而导致的查询检索性能的急剧降低。

（**4**）对现有网络应用的全文检索支持问题

非结构化网络数据库系统不仅能够支持iBASE非结构化数据库的直接上网发布和全文检索，对于传统关系型数据库，如Oracle、SYBASE、SQL Server、DB2、Informix等，也提供了导入和链结的支持能力，用户可以采用导入方式，将传统关系数据库转换为非结构化数据库，再进行网上发布和开发全文检索应用；用户也可采用链结方式，对传统关系数据库构建本地化索引，从而通过本地化索引实现对关系数据库的全文检索支持，iBASE非结构化网络数据库系统充当关系数据库应用服务器，系统的检索效率也将受关系数据库自身检索效率和应用服务器交互效率的影响[6]。

3 非结构化数据库应用分析

网络数据库建设到底采用何种数据库，摆在用户面前的至少有三种方案：关系数据库建设方案、非结构化网络数据库建设方案、关系数据库和非结构化网络数据库共存方案[7]。

在事务处理和数值计算方面，由于关系数据库经过了多年的发展，其在事务处理、数值计算方面具有强大的能力并已被证实。但对于超文本、文档信息管理和数据库全文检索方面，关系数据库通过其MEMO或TEXT字段等也能实现这种信息的存储，而对于这些信息或数据库的全文检索，关系数据库则显得捉襟见肘[8]。

非结构化网络数据库系统则完全解决了网上数据库的全文检索问题，通过其独特的单汉字、单英文词、英文字母的索引方式及树索引算法，能够高效地解决数据库的网上全文检索问题，构造出强大的网上全文搜索引擎。因此，在对于超文本、文档信息管理和数据库全文检索方面，非结构化网络数据库建设方案应为首选。

然而，网络数据库建设并不能从严格意义上按以上两种情况进行区分，但有一点可以肯定的是，大多数网络应用都会有全文检索或构建搜索引擎的需求，从理论上讲，除事务处理能力外，非结构化网络数据库能够处理所有关系数据库支持的网络应用方式，并能对数据库进行全文检索扩展，也就是说，完全可以利用非结构化网络数据库构建独立的网络应用。

对于一些特殊的网络应用，我们也可以采用关系数据库和非结构化网络数据库两者共存的建设方案，实现两者的无缝集成，以发挥两者各自的长处。

主要参考文献：

[1] 孟小峰,周龙骧,王珊.数据库技术发展趋势[J].软件学报,2004,(12)

[2] 向海华.数据库技术发展综述现代情报,2003, .

[3] 王娣.多媒体数据库技术综述情.报杂志,2001,

[4] 吴广印,胡亚莉.非结构化网络数据库在图书情报服务中的应用.图书情报工作,2000,

[5] 阎同喜.数据库技术发展概述机械管理.开发,2004,

[6] 赵淑梅,牛宏霞.新型的数据库技术——XML数据库系统综述.郑州铁路职业技术学院学报,2004,

[7] 陆晔,吉增瑞.数据库系统安全技术综述高性能.计算技术,2001,

[8] 李慧,颜显森.数据库技术发展的新方向——非结构化数据库情.报理论与实践,2001。