Oracle Text 全文检索技术在文档资料管理中的应用①

李培军, 毕于慧, 张 权, 董 玮

(61139 部队信息中心, 北京 100091)

摘 要: 本文利用 Oracle Text 全文检索技术, 根据数据库业务逻辑构建了关键词表, 通过为关键词表建立索引的方式进行检索, 提高了检索效率; 以 ViusalC++6 为开发平台, 采用 C/S 结构技术研发了多类型文档资料管理系统, 实现了办公文档资料的高效管理.

关键词: Oracle Text; 全文检索; 文档资料管理系统

Application of Full-Text Search of Oracle Text in Documents Management

LI Pei-Jun, BI Yu-Hui, ZHANG Quan, DONG Wei

(Information Center of 61139 Army, Beijing 100091, China)

Abstract: Based on the full-text search of Oracle Text, this article first created key words table according to the logical database, the search efficiency was improved used by creating index for the table; and then a documents management system for multi-type files was developed on the platform of Visual C++6 with C/S structure technology to manage official documents efficiently.

Key words: Oracle Text; full-text search; documents management system

1 引言

随着计算机和信息技术的发展, 办公自动化在很多业务部门应用越来越广泛, 许多企业、组织机构的部门每年都产生大量的电子文档资料, 如何高效管理这些文档资料, 方便用户及时快捷地查找自己所需要的信息资源, 是一个亟待解决的问题. 全文检索是以文本数据为处理对象, 根据数据资料的内容来实现信息检索的新一代的信息管理技术. Oracle 数据库提供的 Oracle Text 全文检索技术具有简单易用、成本低廉、功能强大的特点[1], 很适合一般单位实现多种文档格式和复杂搜索标准的文档管理系统的需求. 本文基于 Oracle Text 全文检索技术设计并实现了文档资料管理系统, 在实践中得到了很好的运用.

2 系统总体设计

2.1 系统结构设计

系统通过数据库访问接口层, 以组件的方式将功能进行标准化封装, 不仅提供文档子系统使用, 而且还可为其它子系统调用. 经过封装的接口完成功能的实现, 上层表现框架界面为用户提供可视化操作. 系统支持的文档文件类型包括 Office 文档、图像、图形、声音、影像、演示文稿等类型. 文档资料管理系统总体结构如图 1 所示:

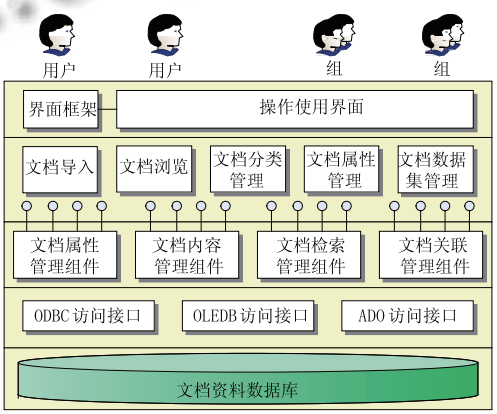

图 1 系统总体结构图

① 收稿时间:2013-08-18;收到修改稿时间:2013-10-28

2.2 系统功能设计

2.2.1 系统维护

系统维护模块主要是对不同的登陆用户的管理以及对数据库数据的管理, 主要包括用户的权限和个人信息的管理、系统设置、数据备份和数据还原等功能.

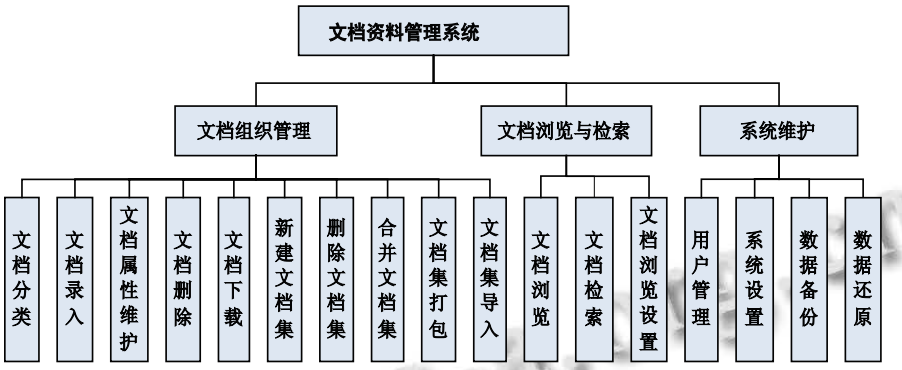

图 2　系统功能结构图

2.2.2 文档组织管理

(1) 文档分类

文档分类主要是对单个或多个文档进行分类, 分类支持层次结构, 例如: 主题→分类→子分类, 主题代表分类标准(例如: 主题为格式, 表示按文档格式分类). 分类层次和分类标准提供用户订制功能以及相应对分类层次和分类标准的维护功能, 包括分类层次和分类标准的增删改.

(2) 文档管理

① 文档属性提取: 自动抽取文档文件属性信息和文档本身固有的属性信息, 并存储在文档文件属性表和文档属性表中. 其中包括通用文档文件属性和专用属性.

② 文档录入: 文档录入提供录入文档到文档数据库的功能, 包括单个文档录入和批量文档录入. 单个文档录入: 选择用户指定的文档, 编辑、录入文档的属性信息, 支持用户默认属性空值, 将文档上传至文档数据库. 批量文档录入: 选择一个以上的用户指定文档, 支持文件夹选择和用户对文档属性自动录入模式及缺省值的设定, 支持同种格式和不同格式文档, 调用文档上传接口, 将文档上传至数据库.

③ 文档属性维护: 录入文档时, 提供对文档内容属性的标注功能; 同时提供对已标注的内容进行编辑和修改的功能. 可标注的文档内容属性分为通用属性和各类型文档的私有属性. 通用属性包括: 所属任务、内容简介、关键词、密级、资源类型等.

④ 文档删除: 文档删除提供删除文档数据库中指定的文档, 包括其内容标注信息和索引信息. 支持单个文档删除和批文档删除, 批文档删除包括不同格式或同种格式.

⑤ 文档下载: 提供用户将文档数据库中的文档下载到指定位置的功能. 用户可直接从文档导航区选择所要下载的文档, 或者通过查询定位所要下载的文档, 下载选定文档.

(3) 文档集管理

文档管理系统可能安装在不同级别用户机器上, 根据工作的需要, 下级用户可能根据需要上报的文档, 上级用户根据需要下发的文档, 因而需要提供文档集管理功能, 完成文档集的新建、删除、插入、合并、生成和录入等操作. 方便各级用户文档的交流.

2.2.3 文档浏览与检索

文档浏览、检索与展现模块在文档组织与管理的基础上提供灵活的文档浏览和查询功能, 同时提供浏览模式的设置.

(1) 文档浏览

文档管理系统提供多种方式的文档浏览功能, 包括基于分类的文档浏览方式、基于入库时间的文档浏览方式和基于格式的文档浏览方式.

(2) 文档检索

文档检索提供基于关键词、文档属性的文档检索功能, 检索模式包括如下几类: 无限制的关键词检索, 基于限定范围的关键词检索: 支持用户选择设定检索

的范围, 如文档分类、文件类型等的基础上对文档库进行关键词的检索. 特定属性的检索: 支持用户对于文档的各种可检索属性进行相应的检索, 具体可通过文档的类型、时间、作者、分类等属性进行检索.

3 实现的关键技术

3.1 基于关键词表的 Oracle Text 全文检索

Oracle 对数据库进行全文检索的原理是对段落性的文本进行逐词分解, 并针对词出现频率, 出现位置进行标记, 按照词本身的编码顺序存储为索引文件. 这样, 在针对关键词进行检索的时候, 就不会遍历所有的文本数据记录, 而是根据索引文件进行有序查找, 对于海量的数据记录而言, 也只需要很少次数的指针跳转即可完成搜索, 而无须完整遍历整个数据表或文件集. 为了进一步提高数据库的全文检索效率, 减少数据遍历的频率, 本文通过构建关键词表与全文检索相结合的方式来实现文档检索. 关键词表是业务逻辑索引表, 它将数据库的诸多表具体化成多个对象(专题), 为利于检索, 将尽可能多的关键词(数据库各表的字段内容)放入关键词表和摘要表内, 通过加字和组合, 形成符合文字表述形式的语句(摘要), 实际上是数据库字段内容到概略知识的组织过程, 并实现关键词与关键表(TableID)及关键行(rowed)[5]的映射, 从而为数据的快速检索提供便利. Oracle 的 varchar2 类型可支持 4000 个字符, 基本可以满足数据需要. 对于 varchar2 字段, 构建索引时间较短, 经过测试, 对于 10 万行数据的关键词表(摘要为 varchar2(4000)), 构建摘要的索引仅需 15 秒; 如果不构建关键词表, 直接对表里的字段建立索引, 当数据库里的表之间的业务逻辑关系越复杂, 建立索引的时间越长.

根据业务逻辑, 通过多表关联得到关键词表所需要的字段数据, 相关插入数据的 SQL 如下所示:

Insert into key_souku_table (KEY_WORD, ABSTRACT,TABLE_ID,ROW_ID,KEY_NM) SELECT a,b,c,d from t1,t2,t3,t4 where …

key_souku_table 是新建的表, 它有五个字段, 根据不同的业务逻辑, 可以插入不同的数据内容. 当把数据库所有的逻辑关系理清后, 就可以将数据库的全部字段内容按照 E-R 逻辑关系组合到关键词表中, 相当于为全数据库建立了一个基于关键词的目录. 用户还可以添加各类统计信息到关键词表以提供基于数据统计分析结果的全文检索.

通过 Oracle 的全文检索功能, 为关键词、摘要及其它大字段（BLOB、CLOB）建立索引, 其 SQL 语句如下所示:

Create index index_souku_abs_content on key_souku_table(ABSTRACT) INDEXTYPE is ctxsys.context parameters('lexermy_lexer');

3.2 查询及显示技术

数据库数据查询及展现的设计, 常常涉及图片、文字、表格、多媒体等元素, 单纯利用 VC 程序实现, 需要进行客户区、非客户区重绘, 维护图片载入、GDI 绘制、消息处理、重载各种控件, 给开发人员带来极大的工作负担, 且收效甚微. 本文将网页技术强大的页面编辑、数据展现、网页特效等功能融入程序中, 将 HTML 与 MFC 相结合, 可以不借助第三方组件做出新鲜生动的显示界面, 查询结果利用 FLASH＋XML 技术进行显示, 具体步骤包括以下几个部分:

(1) 获取用户的查询内容, 通过 HTML 语言的 <FORM> 表单来实现数据查询页面设计, 采用 DOM(文档对象模型)遍历 HTML 中的表单(form)并枚举出表单域的属性, 从而动态提取用户输入, 利用 VC 实现与互联网主流搜索引擎相类似的界面.

(2) 响应网页提交命令. 用户在相应的检索表格中输入需要查询的短语后, 点击提交(搜库一下), VC 通过重载 CHtmlView 类的 OnBeforeNavigate2 函数, 截获网页提交命令, 并交给 VC 程序响应, 获取表单集合内容(即模糊查询内容), 调用数据全文检索函数, 最后将检索结果生成相关的 HTML 语言.

为了实现在网页中能调用 VC 中的函数, 需要重载导航处理器 OnBeforeNavigate2(), 实现“app:”伪协议, 传递“app:”链接到一个虚拟协议处理器. 因为 app: 是假协议, 所以需要设置 pbCancel 参数为“TRUE”, 以停止掉这个导航. 具体代码如下:

```
voidCHtmlCtrl:OnBeforeNavigate2( LPCTST lpszURL,DWORDnFlags,LPCTSTRlpszTargetFrameName, CByteArray&baPostedData, LPCTSTR lpszHeaders, BOOL* pbCancel )
{
    const char APP_PROTOCOL[] = "app:";
    intlen = _tcslen(APP_PROTOCOL);
  if (_tcsnicmp(lpszURL, APP_PROTOCOL,
```

```
len)==0)
    {
        OnAppCmd(lpszURL + len);
        *pbCancel = TRUE;
    }
    }
```

在嵌入的浏览器 CHtmlView 类中定义一个虚函数 OnAppCmd(), 处理“app: ”命令, 当浏览器准备导航到“app:foo”时, 这个函数被调用, 参数 lpszURL 的值为“foo”, 从而实现网页链接调用 VC 中的函数, 通过字符串参数 lpszURL 的分段截取, 还可实现多参数的传入.

(3) 将查询结果按照一定的输出格式显示给用户, 并将标题、关键词和摘要按谷歌或百度的显示样式输出, 根据关键词查询的记录得分实现查询结果的排序, 并提供查询结果的详细链接, 当用户要获取更详细的查询结果, 可根据数据库字段组织逻辑, 结合数据类型决定输出模板风格, 将结果数据整合在一定的模板中, 按照图、文、表形式组织结果网页, 形成图文并茂的查询结果展现给用户.

(4) 查询结果的展现. 对于文字、图表, 可利用 HTML 网页规范显示, 网页具有符合用户查看习惯, 显示方式多样等特点; 对于图片及其说明文字通常利用 FLASH+XML 进行显示, FLASH 具有极强的数据交互和展现能力. 对于普通网页, 可以采用 CFile 或者 CStdioFile 创建后缀名为.html 的文件, 然后写入 html 脚本实现; 而对于 FLASH+XML 模式, 只需修改或重写 XML 文件即可生成数据的动画展现. 最后利用嵌入的浏览器显示生成的网页文件或 FLASH, 从而实现数据的展现.

4 系统实现

采用 C/S 结构, 以 Visual C++6.0 为开发平台进行系统实现, 图 3 是系统的登录界面, 图 4 和图 5 分别是用户登录后的主界面和文档资料导航树, 用户可以根据实际情况按照时间或者内容构造资料导航目录树. 系统设计完成后, 在本单位的多个部门进行了试运行, 能够较好地满足需求.

图 3 系统登录界面图

图 4 系统主界面

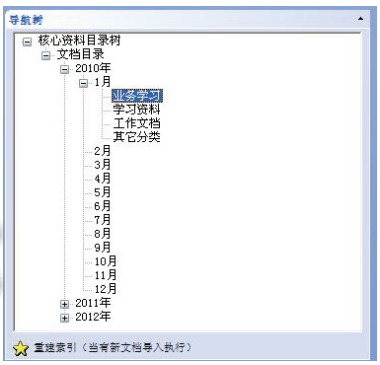

图 5 文档资料导航树

5 结语

本文针对如何更好地管理历年办公电子文档资料问题, 利用Oracle全文检索技术的优点, 采用C/S模式开发了文档资源管理系统, 实现了对各种格式文档的存储、组织管理以及快速浏览查询等功能, 极大地提高了文档的查询和管理效率. 系统经过多次测试, 运行稳定, 效果良好, 具有一定的推广价值.

参考文献

1 朱凡微,吴明晖,金苍宏.基于关键字的数据库搜索研究综述. 计算机应用研究,2008,25(11):3238−3242.

2 Caverlee J, Roccod L. Discovering interesting relationships among deep web database: a source-based approach. World Wide Web, 2006, 9(4): 585−622.

3 李尚初.Oracle 的全文检索技术.哈尔滨师范大学自然科学学报,2009,25(4):92−95.

4 杨应全.Oracle 全文检索技术在高校图书馆的应用.现代情报,2008,9:159−161.

5 赵德玉.Oracle 数据库 rowid 深入探析.广西轻工业,2009,7:76−79.