

基于全信息管理的工程材料数据库系统

王龙义, 李仁旺

(浙江理工大学机控学院, 杭州 310018)

摘要:为实现工程材料数据库的全信息管理、基于全信息管理的材料信息高级查询,尤其是全文匹配搜索,该文分析系统功能,设计基于全信息管理的工程材料数据库系统。重点分析用户动态数据管理与查询算法,提出并实现一种基于单词集提取存储与匹配查询的格式文档全文搜索方法。给出的工程材料数据库系统查询实例表明,该方法具有先进性和实用性。

关键词:工程材料数据库;全信息管理;全文搜索

Project Material Database System Based on Full Information Management

WANG Long-yi, LI Ren-wang

(College of Mechanism and Auto-control, Zhejiang Science and Technology University, Hangzhou 310018)

【Abstract】In order to realize the full information management of the project material database, and the advanced inquiry based on the full information management of material information especially full text search, this paper analyzes the system function, and realizes the project material database system based on the full information management. It analyzes user's dynamic data management and inquiry algorithms especially, puts forward and realizes a kind of full text search based on word collection picking-up, storing and match inquiry form file. The project material databases system provides inquiry instance to indicate advance and practicability of this method.

【Key words】project material database; full information management; full text search

1 概述

在机械产品设计中,材料选择的重要性愈来愈显著,它对机械产品性能、价格、造型、重量等都有直接影响。工程材料数据库是为工程技术人员提供详细材料信息而开发的工程数据库。随着网络技术的发展,工程材料数据库系统从局域网内的C/S结构向面向广域网的B/S结构发展^[1-2]。

现有材料数据库系统在内容上可提供材料的基本信息(牌号、名称、类别等)、物理属性信息(抗拉强度、屈服强度、密度等)、化学特性(各种元素含量、耐腐蚀性等)信息及用户自定义形式数据信息管理^[3-4];在库的组织上,材料数据库已具备良好的开放性与可扩展性,即任何种类材料均可无限制增加其子类材料,并为材料创建其基本数据、属性表或自定义数据格式二维表^[4-5]。

但工程材料数据库系统仍存在一些不足,尤其是缺少全信息管理功能。全信息管理除了对材料的基本信息和属性数据进行管理外,还需要对材料的各种文档信息(包括一些说明或注释性的文档,通常以 Word 的 doc、Acrobat 的 PDF 或文本 txt 格式存在)进行附加、移除、查询等处理操作。在材料数据库的上述文档中搜索符合某个字符串匹配的材料很费时,需要采用先进高效的搜索技术。

全文搜索技术随互联网技术发展而来,是在大量网页中实现字符串匹配搜索的查询技术^[5],它对互联网上各个网站上每个网页的全部内容进行字符串匹配搜索,检索到的结果不是站名、网址或内容提要,而是与输入关键词相关的一个网页地址的一小段,因此,全文搜索算法的关键是信息存储方式及其查询效率。目前全文搜索技术的应用较少,未在材

料数据库系统中得到应用。本文讨论了材料数据库系统全信息管理模式数据库的实现,以及基于全信息模式下实现材料高效查询的全文搜索技术。

2 材料全信息管理的数据库设计

工程材料大致包括如下 5 类信息:

(1)基本信息,包括材料牌号、材料名称、标准代号、所属类别(父类材料)、供货状态、优选级别等。

(2)物理性能,包括密度、抗拉强度、屈服极限、杨氏模量、泊松比、比热、热传导系数、电阻率等。

(3)化学性能,包括常见各种元素(C, Al, Cu, S 等)的含量、耐腐蚀性、耐酸性、抗氧化性能等。

(4)文档类信息,包括以 txt, doc 或 PDF 等各种文档形式存储的说明信息、标准手册信息、图表信息等。

(5)动态属性数据,即每种材料包含的二维表格数据,例如材料规格表、分类与尺寸外形表等。

在本系统中,动态数据表名称以材料牌号开头,根据材料的具体内容,由用户设计字段并输入行值,系统动态生成关于该材料的属性表。

对材料数据库进行数据库设计时必须完整考虑对上述信息的管理,即对上述每类信息进行相应的表格设计。为了实

基金项目:国家自然科学基金资助项目“动静脉相结合制造系统及自组织设计仓库研究”(50675208)浙江省自然科学基金资助项目“无比列技术扩散网络研究”(Y605187)

作者简介:王龙义(1975-),男,硕士,主研方向:网络化设计与制造;李仁旺,教授、博士

收稿日期:2007-10-28 **E-mail:** franckywang@msn.com

现基于全信息，特别是文档类信息的全文搜索，需要增加一个单词集信息表，各表格存储每种材料的文档中所含的字词。材料数据库全信息管理的数据库中的主要的表及其关系如图 1 所示。

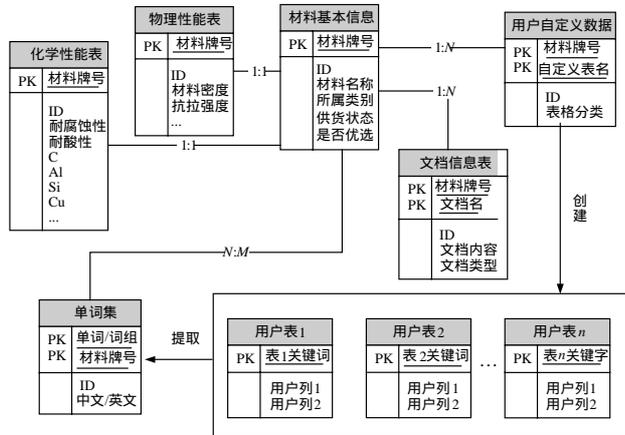


图 1 全信息管理的材料数据库系统 E-R 图

3 全信息管理的实现

基于全信息的工程材料数据库管理包括对材料各种信息的增加、修改、删除及多种查询操作。下文重点介绍与全信息管理相关的几个关键技术的实现。

3.1 用户自定义数据的动态添加

用户自定义数据表格的添加通过输入界面获取用户表名字及表的字段信息(包括各字段名和字段数据类型)。创建 SQL 语句、调用数据库对象 ExecuteSQL 命令建立用户表的伪代码如下：

```
Function CreateUserTable(string sTblName, string sFields[], string
sFieldsType [], int FieldCounts)
{判断表名 sTblName 的合法性;
判断数据库中表 sTblName 是否存在;
判断字段集 sFields 的合法性,至少含有一个字段,否则提示不能创建,并退出
String strSQL="Create Table [" & sTblName & "]"(" & sFields (0)
& sFieldsType(0);
For (int i = 1 To FieldCounts )
strSQL += "," & sFields (i) & sFieldsType (i);
strSQL += ")"
ExecuteSQL(strSQL); //执行 SQL 命令}
```

3.2 材料删除操作

删除某个牌号的材料是较复杂的事务处理过程，其主要步骤如下：

- Step1 开始事务；
- Step2 删除基本信息表中关键词为该牌号名的记录；
- Step3 删除物理属性表中关键词为该牌号名的记录；
- Step4 删除化学属性表中关键词为该牌号名的记录；
- Step5 删除单词集表中关键词为该牌号名的记录；
- Step6 获取用户自定义数据表中关键词为该牌号名的记录，得到用户表名字集合；
- Step7 删除数据库中上述用户表；
- Step8 删除用户自定义数据表中作为该牌号名记录的关键词；
- Step9 提交事务。

事务执行过程中任何阶段出现操作异常或错误，自动滚回事务开始处，以保证数据库中数据的一致性和正确性。

3.3 文档信息添加

文档信息的存储不采用文档路径引用方式，而是通过数据库大字段(image 类型)存储二进制文档来实现，保证了数据库系统的独立性。文档信息的存储过程包括文件压缩、文档数据提取及大字段数据存入等。为了实现全文搜索，对文档进行字符处理，提取文档中的单词集，存入“单词集”表。该算法过程如下：

- Step1 由用户选择文档，得到文档名和类型；
- Step2 开始事务；
- Step3 打开文档，获取文档单词集，存入数据库“单词集”表；
- Step4 调用压缩组件对文档进行压缩，得到压缩文档名；
- Step5 以二进制打开压缩文件，读入内存缓冲区；
- Step6 打开数据库文档信息表，存入文档类型、材料牌号和文档名，将内存缓冲区中的二进制信息存入大字段文档内容中；
- Step7 提交事务。

Step3 为实现后续全文搜索做准备，算法执行过程如图 2 所示。对文档进行单字(英语为单词)提取，若文档为文本文件则直接提取，若为其他格式文件如 doc 或 PDF，则利用 Word 软件或 PDF 软件的 API 接口提出文本部分。对文本进行字符串处理，先提取单字，再对单字进行组合，形成词组集合(不重复)后，将该材料牌号和词组集合存入单词集表中，形成词典以供查询。

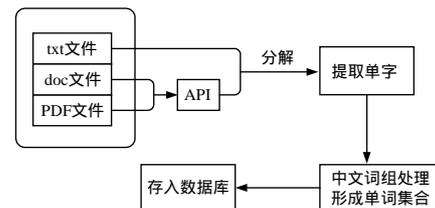


图 2 文档单词集的提取过程

对于汉字词组的处理，采用正向最大匹配算法^[7]，按一定策略将待分析的汉字串与词典进行匹配。采用一种改进的最大匹配分词算法，克服了局部范围最大匹配的局限，体现了长词优先原则，在一定程度上解决了歧义切分问题。

为了提高分词效率，结合全文信息检索中汉字处理的实际需求，对上述算法进行改进^[8]：(1)构造一部基本通用分词词典及停用字表、专业词典和专业同义词典，停用字表中收录了没有检索价值的字符，如标点符号、叹词、介词、助词等虚字；(2)结合不同应用领域背景构造不同专业词典，收录该领域的专业名词和术语；(3)汉语文献作者使用语言的多样性和非规范性导致表达同一概念的关键词常呈现多种字面形式，为了避免词汇差异造成的信息检索不确定性，本文构造了专业同义词典，即使用同义词典实现同义词扩展检索，以解决表达差异问题。

3.4 文档信息显示

在 Web 页面浏览每种材料的全部信息时，对普通信息利用 HTML 超文本实现，对文档类型如 Word 或 PDF 采用 Word 控件来显示 Word 文档，直接利用超级链接可实现 PDF 文档的显示。文档信息显示的主要步骤如下：

- Step1 打开数据库，获取当前材料的文档内容和文档类型；
- Step2 将大字段文档内容放入内存缓冲区；

Step3 将缓冲区的内容存入临时文件；

Step4 解压缩临时文件，根据文档类型给定临时文档名和扩展名；

Step5 根据文档类型调用相应显示控件，将解压缩的文档装载显示。

3.5 文档全文搜索

全文搜索是本系统实现材料查询的一种高级搜索技术，其基本思路是在材料录入时，特别是文档录入时进行处理，将文档信息提取存入单词集——相当于字典中。执行查询时全文搜索只是在此表中进行搜索，而非在文档中进行适时查找，从而极大提高了全文搜索效率。算法基本思想如图 3 所示，先根据用户输入对输入字符串进行词组划分；再根据文档文件(可能是文本文件、Word 文件或 PDF 文件，如果不是文本文件，还需要调用相应 API 函数进行处理)提取的单词集进行搜索；然后将每个字与材料名称进行组合，形成一个键值对。当所有数据文件都被处理完之后，就会形成一个以该字作为关键词的映射表，返回不重复的材料牌号。

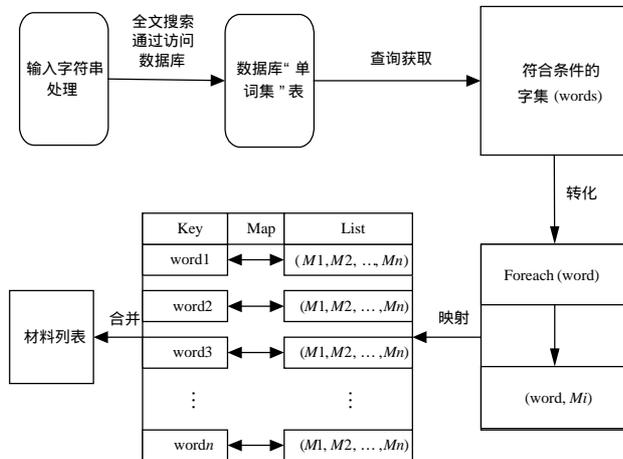


图 3 全文搜索的实现

4 基于 Web 的工程材料数据库系统运行实例

材料数据库浏览包括左边的树控件和右边的显示网页。操作过程与帮助系统相似，点击树控件上的节点，展开其子节点，并在右边显示其子项或具体内容。材料浏览页面左面部分显示选中材料包含的基本信息，一般包括基本属性和属性表；右面部分是基本信息的详细信息，即材料的基本属性，包括材料名称、中文名称、手册优选、企业优选信息。材料属性表通常是同一大类材料的详细信息。

在材料查询中，可以根据不同条件查询到所需材料。先选择查询范围，查询范围越小，获得的材料越少。查询条件分为基本选项和高级选项，基本选项是根据材料的基本属性条件查询，例如材料名称、中文名称等；高级选项是根据材料的专有属性条件查询，自由条件查询是在文档信息中进行全文搜索。选择需要查询的选项并点击“添加”，条件会被添加到查询列表中。如果不需要某个查询条件，可以在列表中选中它，点击删除“查询条件”或“清除查询条件”。完成所有条件的添加后，点击“查询”，查询结果被显示，点击其中的项，便可浏览其基本信息。点击“点子手册”可显示手册的文档信息，如图 4 所示。

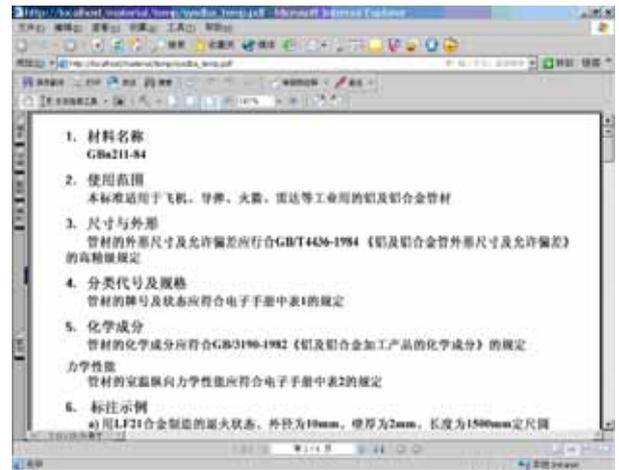


图 4 文档信息显示界面

5 结束语

本文工程材料数据库系统采用目前最先进的数据库系统技术、计算机网络技术和系统实现技术开发而成，系统界面友好、使用方便、运行稳定、查询快捷。与国内其他材料数据库系统相比，具有如下优点：(1)界面美观。本系统采用最新的 WebControls 树形控件，直观地显示材料数据库的类别及层次关系；(2)数据库设计采用动态模式，允许无限级别的材料添加。(3)全信息管理。在材料数据库中，可以浏览材料的书册 PDF 图片信息，并得到其相关文本或表格信息，便于用户引用和拷贝。(4)全文搜索功能。这是本系统最具特色的功能，基于全信息管理，系统提供了满足一定物理性能(如拉伸强度大于 100 MPa)、化学成分(如含 C 量大于 0.5%)或其他任意自定义条件的组合查询。

本文材料数据库管理系统功能强大、使用方便、开放性好，已在银河科技集团公司内部运行，并显著提高了该公司产品的设计效率。

参考文献

- [1] Faraj M I, Alshawi G, Aouad T, et al. An Industry Foundation Classes Web-based Collaborative Construction Computer Environment: WISPER[J]. Automation in Construction, 2000, 10(10): 79-99.
- [2] 陈扬枝, 张见威. 基于 Web 的机械设计材料数据库系统[J]. 现代制造工程, 2001, 5(12): 16-17.
- [3] 李浩, 朱林. 面向服务架构的工程材料数据库系统开发[J]. 计算机与数字工程, 2006, 34(11): 150-153.
- [4] 谈小康, 吴义忠, 王飞. 基于 Web 的工程材料数据库系统的设计及开发[J]. CAD/CAM 与制造业信息化, 2003, 2(3): 98-101.
- [5] 马骏. 基于 Web 技术的工程材料数据库开发与研究[D]. 武汉: 华中科技大学, 2004.
- [6] 陈康, 许婷, 戴文俊, 等. 基于 Web 的全文搜索引擎的设计与实现[J]. 计算机工程, 2005, 31(20): 51-53.
- [7] 唐培丽, 胡明, 解飞, 等. 全文检索搜索引擎中文信息处理技术研究[J]. 情报科学, 2006, 24(6): 895-900.
- [8] 郭辉, 苏中义, 王文, 等. 一种改进的 MM 分词算法[J]. 微型电脑应用, 2002, 18(1): 13-15.