基于分类标引的文件管理系统①

徐宏发[1]，陈 蔚[2]

[1](中国人民公安大学 公安情报学系，北京 100038)
[2](北京培黎职业学院 国际商务系，北京 100085)

摘 要： 随着海量公开信息源的涌现，大容量存储设备进入个人领域，个人拥有庞大的信息已成为现实。与传统的图书信息不同，个人信息源是根据个人的爱好和需要所构建，因此很难像图书一样制定《中图分类法》来管理个人信息，大量的信息搁置在个人的存储设备上，我们已经无法用传统的手段进行有效管理了，比如目录管理。因此，如何有效地管理个人信息成为当前信息管理领域的一个重大挑战。本文提出基于分类标引的思想构建文件管理系统来实现对个人计算机的信息管理，主要内容包括系统的体系构架、功能设计等，以及面临的挑战。

关键词： 个人信息管理；分类标引；文件管理系统；自学习；分类体系库

File Management System Based on Classifying Indexing

XU Hong-Fa[1], CHEN Wei[2]

[1](Information Department, Chinese People Public Security University, Beijing 100038, China)
[2](Department of International Business, Peking Bailie University, Beijing 100085, China)

Abstract: With massive open information resource welling up and large-capacity devices entering into private life, it come in true that everybody own gigabytes information. On contrast with traditional books, personal information resource derive from building by their hobby and demand, then it is difficult to manage personal information as managing books with the classification of chinese books. Since large information landed on personal storage devices, we can't manage them efficiency by such as directories. It is a big challenge to manage personal information more efficiency in information management. The paper builds up file management system based classifying indexing to handle information management of personal computer, and introduces architecture, design and practice, also challenges of the file system.

Keywords: personal information management; classifying indexing; file management system; self-learning; classifying database

随着大容量存储设备进入个人办公、家庭娱乐等领域。网络环境下的海量信息的有序组织与管理的问题同样摆在个人面前。如何有效地管理个人信息(包括数据、文件、目录、连接等)成为个人信息管理[1,2]领域的一项重要课题。本文主要探讨个人计算机中的目录和文件的组织与管理问题。

当前个人计算机的信息管理主要是文件管理，基本上是采取目录管理的方式。随着文件的增多，目录管理的方式越来越难以满足的应用需要，比如，一个文件 a 可能属于 A 类，那么，可以创建一个目录 A，将文件 a 放在目录 A 下，但是，如果文件 a 同时属于 A 类、B 类和 C 类，那么该如何存放该文件呢，是否在目录 A、B、C 下各自都要存放一个 a 文件呢，显然，这样的解决方案面临严重的数据冗余，难以删除，难以更新，难以维护。那么将这些文件数据放到数据库中，可以么？当然可以，但是，如果每次访问文件，

① 收稿时间:2010-06-02;收到修改稿时间:2010-07-11

都要先打开数据库，这不是大家愿意看到的。可不可将目录管理文件的优势同数据库存取数据的优势相结合来解决这个问题呢？当然是可以的。

另外一个问题是，目前文件的固定属性满足不了我们对文件组织和管理的需求。以 ABCD.Doc 文档为例，在 windows xp 操作系统中，定义了文件类型、打开方式；位置、大小、占用空间；创建时间、修改时间、访问时间；操作权限等属性。尽管操作系统提供了强大的桌面搜索引擎，但是对于用户而言，大多数人仍是采用层次目录来组织并浏览文件。也就说，当操作系统的预定义属性满足不了我们对个人信息的组织、管理和检索、利用时，如何来解决这些问题呢？显然，统一的固定的分类体系并不适合个人信息，比如《国图分类法》，因此，需要研究用户的信息管理行为，构建一种可以自适应的分类体系，根据用户的习惯构建动态的内容分类体系无疑是我们应该努力的方向，并且应考虑将分类体系植入文件或目录本身的固有属性中。举例来说，如何实现保存期限的设置，用户可以根据自身的需求给文件提供一个期限属性，该属性包括如下可选值：永久保存、有期限保存、临时保存等类别，并且可以给临时保存给予一个固定值，当文件失效日期达到时，系统自动将文件放进垃圾桶中，实现文件的自动删除。

本文提出的基于分类标引(标引的本质是，实现目录文件真实地址同分类体系的映射或关联)的文件管理系统。该系统并不关心目录和文件的真实地址(目录和文件的真实地址由系统自动分配，类似于 DB2 中的自动存储管理，简化系统的存储管理)，只需对目录和文件进行标引，建立分类体系同目录、文件之间的映射。该系统采取分类标引[3]的思想，借鉴数据库系统的分层结构[4]的理念，将系统划分为三层：目录文件管理层、控制层和分类体系管理层，在层与层之间采取映像模式实现逻辑的独立性。具体而言，分类体系管理层实现对分类体系库的管理、分类体系的自学习等；控制层实现组织、管理、更新文件目录与分类体系的映射库等操作；目录文件管理层实现目录文件的增加、修改和删除等功能。分类体系管理层与控制层的映像、目录文件管理层与控制层的映像，分别实现分类体系管理层与控制层、目录文件管理层与控制层的通信。

1 系统架构

本系统采用三层体系结构架构，分为目录文件管理层、控制层和分类体系管理层。同时采用二级映像具体实现层间的通信协议提高各层的灵活性。系统体系结构见图 1。

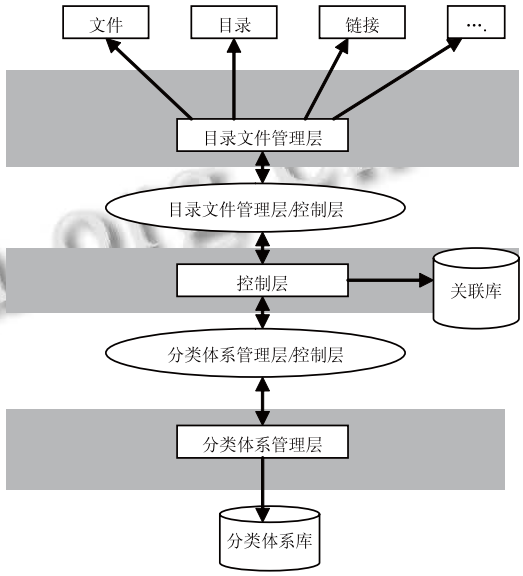

图 1 基于分类标引的文件管理系统

目录文件管理层，位于目录文件与目录文件管理层/控制层映像之间，其主要功能包括：①监控目录文件的位置变化(删除、增加、移动等)，并将操作请求发送给控制层；②接收控制层的指令，实现对未标引文件和标引文件的管理，比如(删除、增加、移动等)。

控制层，位于目录文件管理层/控制层映像与分类体系管理层/控制层映像之间，其主要功能包括：①接收目录文件管理层发送的请求，查询或更新关联库，并反馈请求结果；②接收分类体系管理层发送的请求，查询或更新关联库，并反馈请求结果；

分类体系管理层，位于分类体系库与分类体系管理层/控制层映像之间，其主要功能包括：①分类体系的透明存储；②实现分类体系的更新；③分类体系的自动学习[5,6]；④接收控制层指令，查询分类体系库；⑤向控制层发送分类体系更新请求。

目录文件管理层/控制层映像，其主要功能包括：①实现目录文件管理层与控制层之间的通信；②实现系统可扩展性，如本地通信向分布式式系统的扩展，可以将目录文件管理层和目录文件管理层/控制层映

像部署在本地，将控制层、分类体系管理层部署在远程服务器上；③降低目录文件管理层与控制层之间的耦合度，即只要目录文件管理层/控制层映像的接口不发生改变，目录文件管理层、控制层内部的逻辑代码是可以根据业务进行调整，而不会影响层与层之间的通信接口。

分类体系管理层/控制层映像，其主要功能：①实现分类体系管理层与控制层映像之间的通信；②实现系统可扩展性；③实现业务的可扩展性，引入分类体系管理层/控制层映像提高了业务的可扩展性，即原来是个人的分类体系库，现在可以像“云计算”一样集成海量的个人分类体系库，实现海量分类体系的自学习，进而可以根据人群的分类实现分类体系的标准库，并提供用户可订制的标引库，这点尤为重要，它可以从本质上解决个人信息管理问题(个人信息管理的面临难题本质上是个人分类体系的复杂性和随意性)。④降低分类体系管理层与控制层之间的耦合度。

2　系统设计与实现

根据文件管理系统架构，对需求进行细化，进行系统设计与实现。

2.1 系统数据库设计

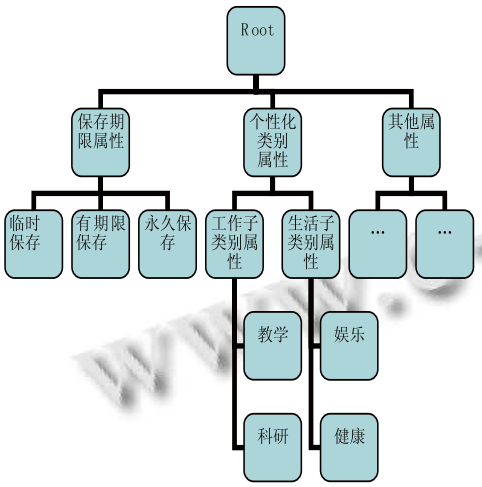

图 2　类别体系示例图

在文件管理系统中，主要存在如下业务实体：目录、文件、分类属性、联系。

文件是标引的实体对象，目录是文件的聚合对象。实体对象，意味着只要对其进行标引，必然会对应确切的分类；聚合对象，是一个准集合的概念，引入目录标引，主要是为了操作上的方便，比如，可以标引一个目录，进而标引整个目录下的子目录和文件；联系是分类体系与目录文件的联系，一个联系的主键是具体目录文件外码和分类属性外码的联合码。建立集合表述：

文件集合 F = {f1,f2,f3,…,fn}

目录集合 D= {d1,d2,d3,…,dm}

分类属性集合 C= {c1,c2,c3,…,ck}，分类属性经过组织，形成具体的分类体系，见图 2。

联系集合 R= {r1,r2,r3,…,ri}

(1) 目录和文件之间是一对多的关系，即一个目录对应多个文件

(2) 文件和分类属性之间联系是多对多的关系，即一个文件可以拥有多个分类属性，一个分类属性可以标引多个文件

(3) 目录和目录之间的关系一对多的关系

(4) 目录和分类属性之间不存在必然的关系，其关系主要体现在目录下的文件同分类属性之间的关系，因此，可以说目录和分类体系之间是多对对关系

(5) 分类属性之间是一个层次关系，是严格的一对多关系，及任何节点仅有一个父节点(除根节点外)

系统的实体-联系图，见图 3，具体数据表设计略。

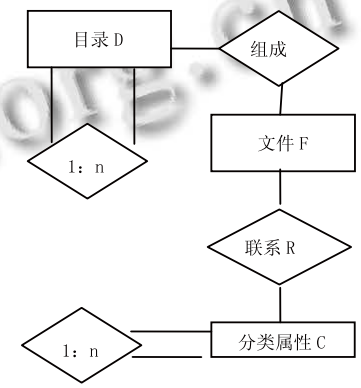

图 3　系统实体关系图

2.2 系统功能设计

系统功能设计主要对系统构架中的功能进行具体化。主要包括目录文件管理层、控制层和标引管理层，以及两级映象的具体功能说明。由于在系统构架一部分已经有了一些说明，在这里只是对重点功能进行细

化。

(1) 目录文件管理层。目录文件管理层应管理目录文件标引初始化、目录文件删除、移动及增加行为。

(2) 控制层。控制层负责维护文件与分类体系库、目录与分类体系库的联系，实现对目录文件管理层和分类体系管理层发生的请求进行处理，并发出指令。

(3) 分类体系管理层。分类体系管理层重要的任务是建立、维护分类体系库等。

① 分类体系库的建立

可以采用两种方式建立分类体系库。一是，借用图书馆学的方法，拟定文件分类体系库。二是，采用分类体系自学习功能，即不手动建立分类体系库。观察和统计用户在使用文件或目录时行为，归纳形成相对成熟的分类体系，在此基础之上，采用一定的规则约束和推理算法，自动构建分类体系库。在实际系统设计和开发时，我们采用两者结合的方法。

② 分类体系库的维护。分类体系库的维护包括分类体系的增加、删除和修改等。

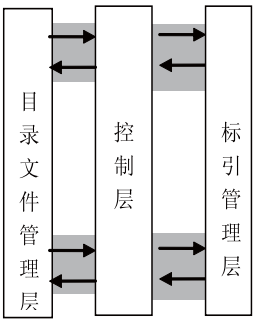

图 4 系统通信主要流程图

(4) 两级映像

两级映像主要是制定和实现通信协议。系统通信主要流程图如下，其中阴影部分是包括二级影响里面的，包括四组基本的通信模式。

① 目录文件管理层向控制层发送请求，控制层处理请求，并反馈。

② 控制层向目录文件管理层发送指令，目录文件管理层响应指令后，并反馈。

③ 分类体系管理层向控制层发送请求，控制层处理请求，并反馈。

④ 控制层向分类体系管理层发送指令，分类体系管理层响应指令后，并反馈。

2.3 举例说明

现以电影文件——“泰坦尼克号.avi”为例，参照图 2 的分类体系模板，来说明本系统与传统的目录管理的区别。在传统的目录管理中，我们会创建一个目录 movie/love 或者 movie/classic 或者 movie/english 等等，但是无论如何创建这些目录，我们都会面临一个问题，到底是创建多个目录，然后将文件“泰坦尼克号.avi”放置其中？还是我们只选择创建一个目录，丢弃其他目录呢？如果选择前者，当我们需要删除或者更新文件时，会面临同步问题；如果选择后者，我们就会放弃浏览时的便捷性以及用户个性化分类。当然传统的目录远不只存在这些局限性，比如如果我想将“泰坦尼克号.avi”的有效期设置为“有限期保存——10 年”，那么在现有的目录管理体系中，我们如何解决呢？

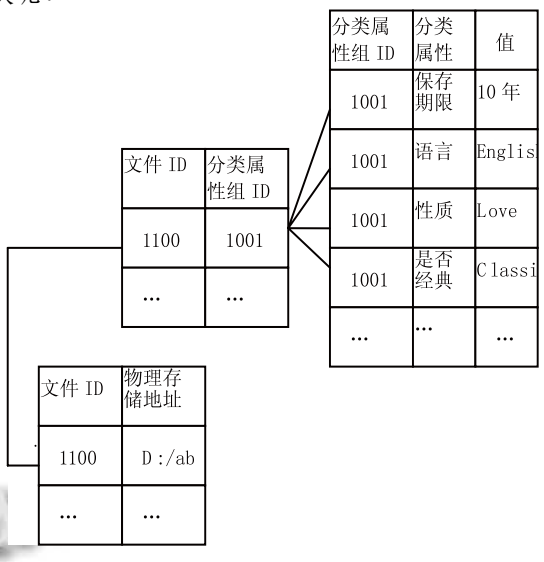

图 5 信息模型关联图

下面，我们看看在本系统中，我们如何解决这些问题。目录文件的存储由目录文件管理层，采用类似 IBM DB2 数据库中的自动存储管理的方法，即用户并不需要知道目录文件的真实的物理存储地址。那么用户是如何浏览目录文件的呢？控制层首先将目录文件分成两大类，一类是标引了的，另一类是未标引的。对于标引的目录文件，将根据分类体系构建一个虚拟的目录层次，供用户从多个角度进行浏览。比如对于文件“泰坦尼克号.avi”，其物理存储地址只有一处，但是可以通过控制层，建立分类体

系同文件的联系，将联系存储至管理库，关联库中存储的内容大致如下：文件 ID | 分类属性组 ID ；分类属性组 ID | 子分类属性 | 可用值。系统总体的信息模型关联图，见图 5。

2.4 难点

(1) 获取文件或目录的变化。笔者采用 DLL 注入的方法，将监控文件或目录的逻辑代码封装成 DLL 注入到 Windows 的系统文件。

(2) 实现分类体系库的自学习功能。包括两个问题，一是在单机环境下用户分类体系库自我的学习，二是在分布式环境下，用户分类体系库之间的学习借鉴。在单机环境下用户分类体系库的自我学习实际上是对用户历史标引分类行为的归纳，同时设置一定的条件加以演绎，形成新的分类体系，这是一种不断的学习和总结的过程。在分布式环境下，除了单机环境下的分类体系库学习之外，更加重视用户之间的标引学习行为，特别同群用户之间的学习，当然，这是基于个人对自身的定位，比如，你将自己定义为教师，或者更加小的分类：高校教师、中学教师、小学教师等。这是本系统以后的努力目标。

对于海量信息的管理成为我们每个个体无法回避的问题，因此，研究有效地管理方式是当务之急。本文提出基于分类标引的文件管理系统只是一个小尝试，希望在此基础上，能够构建一个分布式的自学习分类体系库，形成一个基于个体角色的、类似于《中国分类法》的分类体系，用于指导个人信息的管理，提高个人信息管理的效率和水平。

参考文献

1 Bergman O, Boardman R, Gwizdka J, Jones W. Personal information management. Proc. of the CHI 2004. ACM SIGCHI Special Interest Group. New York: ACM Press, 2004. 24－29.

2 Teevan J, Jones W, Bederson BB. Personal information management. Communications of the ACM, 2006,49:40－43.

3 陆小辉,周金付.共享环境下文献分类标引的一致性.图书馆杂志,2005,7:42－43.

4 王姗,陈红.数据库系统原理教程.北京:清华大学出版社,1998. 22－25.

5 Nguyen DH, Widrow B. Neural networks for self-learning control systems. International Journal of Control, 1991, 54(6):1439－1451.

6 王青,祝世虎,董朝阳,陈宗基.自学习智能决策支持系统.系统仿真学报,2006,18(4):924－926.

(上接第 233 页)

识,它定性定量地分析了上下层节点的各故障模式间的关系。利用 PDM 中结构树、FMEA 信息构建产品数据模型 。从产品数据模型中抽取故障知识建立产品故障树，根据故障的权重来进行检查，直到找到正确的故障原因，当然维护信息也可以同理建立相同的维护树，可以及时的进行维护，保证 ATM 机一直处于良好的工作状态。

4 结论

本文提出了 PDM 结合 FMEA 的对金融故障诊断及维护方法，充分利用金融产品从设计到生产相关文档齐全，各零部队参数详细，有着明确统一的标准的特点及金融产品设计阶段的诊断知识，生成故障树。根据故障树可以预先判明金融设备的哪部分需要维护或者是故障的原因，大大提高了工作效率和诊断精度，并以 ATM 机系统为例进行了说明，验证了其可行性。

参考文献

1 陈勇飞. FMEA 简介.机械工程师,2002,(1):31－32.

2 吴含前,姜澄宇,王宁生.PDM 技术的发展.机械设计,2000,(12):2－3.

3 费胜巍,孙宇,张登峰,等.由产品设计知识生产故障诊断与维护知识的方法.机械设计,2006,(2):7－9.

4 孙宇,彭强,张晓阳,等.基于混合结构树的故障诊断技术研究.计算机集成制造系统,2005,(7):1031－1033.

5 费胜巍,孙宇,张晓阳,等.基于产品结构树和 FMEA 的故障诊断方法研究,2006,(7):238－240.

6 艾民,张利伟.自动柜员机的构造及维护.宁夏科技,2003,(3):32.

7 姜英武.自动取款机故障诊断系统 ATMDES1.0 的设计和实现[硕士学位论文].长春:吉林大学,2006.

8 Kalagnanam J. A system for automated mapping of bill-of-materials part numbers. Proc. of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. 805－810.