(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2009 年2 月19 日 (19.02.2009)

PCT

(10) 国际公布号
WO 2009/021465 A1

(51) 国际专利分类号:
*G06F 17/30* (2006.01)

(21) 国际申请号: PCT/CN2008/072010

(22) 国际申请日: 2008 年8 月15 日 (15.08.2008)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
200710120325.3
2007 年8 月15 日 (15.08.2007) CN

(71) 申请人 (对除美国外的所有指定国): **北京书生国际信息技术有限公司 (SURSEN CORP.)** [CN/CN]; 中国北京市海淀区紫竹院路81号北方地产大厦5层, Beijing 100089 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **王东临 (WANG, Donglin)** [CN/CN]; 中国北京市海淀区紫竹院路81号北方地产大厦5层, Beijing 100089 (CN)。 **姜海峰 (JIANG, Haifeng)** [CN/CN]; 中国北京市海淀区紫竹院路81号北方地产大厦5层, Beijing 100089 (CN)。 **刘宁胜 (LIU, Ningsheng)** [CN/CN]; 中国北京市海淀区紫竹院路81号北方地产大厦5层, Beijing 100089 (CN)。 **王立伟 (WANG, Liwei)** [CN/CN]; 中国北京市海淀区紫竹院路81号北方地产大厦5层, Beijing 100089 (CN)。

(74) 代理人: **北京德琦知识产权代理有限公司 (DEQI INTELLECTUAL PROPERTY LAW CORPORATION)**; 中国北京市海淀区知春路1号学院国际大厦7层, Beijing 100083 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY,

[见续页]

(54) **Title:** METHOD AND DEVICE FOR STORING DOCUMENT DATA IN DOCUMENT BANK SYSTEM

(54) **发明名称:** 文档库系统中存储文档数据的方法及装置

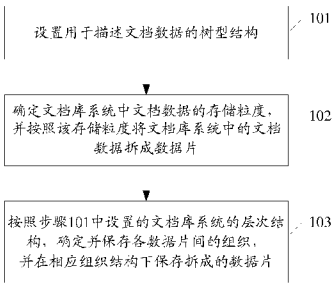



图 1 / Fig1

101 SETTING TREE STRUCTURE THAT DESCRIBES DOCUMENT DATA
102 DETERMINING STORAGE GRANULARITY OF DOCUMENT DATA IN DOCUMENT BANK SYSTEM, AND SPLITTING DOCUMENT DATA IN DOCUMENT BANK SYSTEM INTO DATA PATCHES ACCORDING TO STORAGE GRANULARITY
103 ACCORDING TO SETING HIERARCHICAL STRUCTURE OF DOCUMENT BANK SYSTEM IN STEP 101, DETERMINING AND STORING ORANIZATION BETWEEN EVERY DATA PATCHES, AND STORING SPLITTED DATA PATCHES UNDER CORRESPONDING ORANIZATION STRUCTURE

(57) **Abstract:** A method and a device for storing document data in document bank system are provided, wherein the method includes: the document data is described by a tree structure, when the document data is stored, the document data is split into a plurality of data patches according to the tree structure; a plurality of data patches are stored in the manner of the tree structure, each of the data patches corresponds to a node of the tree structure.

(57) **摘要:**

本发明公开了一种文档库系统中存储文档数据的方法及装置，其中方法包括：所述文档数据按树型结构描述，在存储所述文档数据时，将所述文档数据按所述树型结构拆成多个数据片；以所述树型结构存储所述多个数据片，所述每一个数据片对应所述树型结构中的一个结点。

WO 2009/021465 A1

MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

**(84) 指定国**（除另有指明，要求每一种可提供的地区保护）：ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

**本国际公布：**

— 包括国际检索报告。

# 文档库系统中存储文档数据的方法及装置

## 技术领域

本发明涉及文档存储的技术，特别涉及文档库系统中存储文档数据的方法及装置。

## 背景技术

文档库系统是一种复杂的软件系统，提供大量电子文档的组织、管理、安全、存储和展示等功能。在现有的文档库系统中，文档数据的存储方式为：将所有的文档数据作为一个文件进行存储。通过访问文件来实现对数据的更新和查询。但是，随着信息量日益增加，文档库系统中保存的文档数据量也急剧增多。由于这些海量数据均作为一个文件进行存储，因此导致其数据的更新和查询变得非常复杂，不能迅速定位用户感兴趣的内容，使内容抽取效率较低。可见，目前的存储方式已不能适应日益增多的海量数据。

## 发明内容

本发明实施例提供了一种文档库系统中存储文档数据的方法及装置，能够实现对文档内容进行灵活高效地检索与内容抽取。

本发明实施例提供的文档库系统中存储文档数据的方法包括：

所述文档数据按树型结构描述，在存储所述文档数据时，

将所述文档数据按所述树型结构拆成多个数据片；

以所述树型结构存储所述多个数据片，所述每一个数据片对应所述树型结构中的一个结点。

本发明实施例提供的文档库系统中存储文档数据的装置包括：

第一单元，用于将所述文档数据按所述树型结构拆成多个数据片；

第二单元，用于以所述树型结构存储所述多个数据片，所述第一单元拆分的每一个数据片对应所述树型结构中的一个结点。

本发明实施例提供的文档库系统中存储文档数据的方法及装置中，按照文档数据的树型结构将文档数据拆成多个数据片，然后将该拆分的数据片按照树型结构存储起来。这样将文档数据拆成更细的数据片并保存，当用户检索或访问文档数据时，可以快速、准确定位到数据片的级别，进行灵活高效地数据检索与内容抽取，对于基于网络的文档应用，降低了文档传输过程中的负担。

## 附图说明

图 1 为本发明实施例提供的文档库系统中文档数据存储方法的流程图。

图 2 为本发明实施例提供实现文档库系统中文档数据存储的装置结构示意图。

## 具体实施方式

为使本发明实施例的目的、技术手段和优点更加清楚明白，以下结合附图对本发明实施例作进一步详细说明。

本发明实施例提供的文档库系统中文档数据的存储方法为：将文档数据拆成多个数据片，并记录这多个数据片之间的关系。当文档数据按树型结构描述时，在存储所述文档数据时，将所述树型结构拆分出多个子树，每个子树形成一个数据片，并记录拆分后剩下的树型结构。

具体来说，可以从文档数据树型结构中任意一结点开始拆分数据片，拆分出的数据片是以该结点为根结点的子树，如果该结点为叶子结点的话，该子树就只有一个结点。重复上述过程，就拆出了多个数据片。

拆分点的选择可以是任意的，拆分过程也可以在任意时候终止，拆分出的子树还可以按照同样的方法继续拆分。一种优选方案是为整个树型结构规定一个存储粒度，按照该存储粒度拆分整个树型结构。例如以页为存储粒度时，就从所有页结点开始拆分，即将所有以页结点为根结点的子树都拆分出来。另一种优选方案是对树型结构的不同部分规定不同的存储粒度，按照各自的存储粒

度拆分树型结构中的不同部分。例如某个文档集下的所有数据都从页结点开始拆分，其它数据都从文档结点开始拆分。

图 1 为本发明实施例提供的在文档库系统中文档数据存储方法的流程图。如图 1 所示，该方法包括:

步骤 101: 利用树型结构来描述文档库系统以及文档数据。其中描述文档库系统的树型结构由根到叶子结点依次可以包括下述部分: 文档集、文档、页、层、对象组和版面对象。描述文档数据的树型结构由根到叶子结点依次可以包括: 文档、页、层、对象组和版面对象。这里描述文档数据的树型结构又可称为文档数据的组织结构。

步骤 102: 确定文档库系统中文档数据的存储粒度，并按照该存储粒度将文档库系统中的文档数据拆成数据片。

这里文档数据的存储粒度是与文档库系统所支持的树型结构、以及文档数据本身的组织结构信息相关。

若文档数据的树型结构为文档以及页，那么文档数据的存储粒度可以为文档，也可以为页。若文档数据本身的组织结构只有一层，即文档，不能再拆成页，那么文档数据的存储粒度即为文档。但无论怎样，文档库系统中的文档数据本身的组织结构至少支持文档级，文档库系统的树型结构至少为文档集级。

确定文档数据的存储粒度后，将该存储粒度作为最小拆分单元来将文档数据拆成数据片。

步骤 103: 按照文档数据的组织结构，保存拆成的数据片。

因此，与现有技术中将文档库系统中的所有文档数据作为一个文件的存储方式相比，依照本发明的方式进行文档存储，其存储粒度一定更细，也就能够降低更新、检索的复杂性，进行文档数据的检索和内容抽取时，能够快速定位到存储粒度，从而实现灵活高效的检索和内容抽取。

具体地，在本发明实施例提供的文档库系统中，文档库系统的树型结构可以通过三种形式表示: 可扩展标识语言（XML）数据库方式、文件系统方

式和关系型数据库方式。下面通过三个具体实施例说明上述三种方式的具体实施方式。

在本发明第一实施例中，通过 XML 数据库方式表示文档库系统的树型结构。在该种方式下，本发明实施例提供的文档库系统中文档数据的存储方法流程包括：

步骤 201：设置用于描述文档数据的树型结构。

本步骤中，设置的文档库系统树型结构由根到叶子结点依次可以包括下述部分：文档集、文档、页、层、对象组和版面对象。

表示上述文档库系统树型结构的方式可以为：

利用 XML 数据库的树型结构表示文档库的树型结构，且将文档库中文档级以上的层次映射到 XML 数据库中的集合（Collection）级。

例如，对文档库设置文档集、文档、页、层和版面对象，可以利用 XML 数据库的树型结构表示上述层次，即文档集中包括多个文档，文档中包括多个页，页中包括多个层，层中包括多个版面对象。

步骤 202：确定文档库系统中文档数据的存储粒度，并根据该存储粒度将文档库系统中的文档数据拆成数据片。

本实施例中，确定文档数据的存储粒度可以为：依据文档库系统设置的存储粒度结合文档数据本身的组织结构，确定存储粒度。其中，文档库系统设置存储粒度的方式可以包括：为文档库系统设置固定的存储粒度，和/或在对文档库系统进行参数配置时配置所述存储粒度，和/或在安装文档库系统时指定存储粒度。文档数据本身提供的结构信息可以是如文档是否分页，页是否分层等信息。另外，确定文档数据的存储粒度还可以通过用户指定的方式进行。最终确定的存储粒度可以是该文档库系统树型结构中的任意一级，如页、层等，优选地，确定页为存储粒度。对于文档数据的不同部分，可以按照相同的存储粒度进行拆分，也可以为文档数据不同部分设置不同的存储粒度。

将文档数据拆成数据片的具体方式可以为：对文档数据中的存储粒度相

关数据，以确定的存储粒度为单位进行分解，得到数据片。进一步地，对于文档数据中的存储粒度无关数据，如文档库、文档集的元数据、角色数据、权限数据、导航数据、源文件数据和字库数据等，首先确定该数据在所述树型结构中的对应级，然后将每一级下的存储粒度无关数据拆分成一个或多个数据片，作为对应级中的一个或多个数据片。具体地，可以按照存储粒度无关数据的类别来拆分存储粒度无关数据，例如将导航数据作为一个数据片、将权限数据作为一个数据片等。其中，具体将存储粒度无关数据作为一个数据片还是多个数据片根据实际应用确定。对于属于同一类别的存储粒度无关数据，也可以进一步根据其自身特性拆分成多个数据片。例如，当存储粒度为文档时，存储粒度相关数据则为该文档中的内容数据等，存储粒度无关数据则为文档共享的字库数据、源文件数据、导航数据等。对于字库数据和源文件数据等存储粒度无关数据，可以将不同文档的源文件数据和字库数据拆分成不同的数据片，而不将所有文档的源文件数据和字库数据作为一个数据片；但对于导航数据等存储粒度无关数据，则可以将所有文档的导航数据作为一个数据片。

步骤 203：保存拆成的数据片间的组织结构。

在本实施例中，文档库系统的树型结构是利用 XML 数据库的树型结构表示的。在这种表示方式下，保存数据片间组织结构的方式具体为：将该组织结构映射到 XML 数据库的树型结构中，优选地，可以利用集合保存该映射后的树型结构。

步骤 204：保存拆成的数据片。

本步骤中，保存数据片的方式具体可以为：将数据片转换为 XML 文档加入到对应的集合中。或者，还可以将数据片作为文件系统中的文件或关系型数据库中的二进制大对象保存到对应的集合中。

至此，本实施例提供的文档库系统中文档数据的存储方法流程结束。在应用上述方法后，可利用 XML 数据库的管理功能来管理文档库系统中的文档数据，完成应用程序的请求，定位相应数据并返回给应用程序。

使用 XML 数据库方式最大的好处是抽取信息时可以获得最大的灵活性，通过 XML 数据库支持的 X 路径（XPath）及 X 查询（XQuery）等方式，应用程序可以方便地访问到文档对象、页对象、层对象、层内对象（如图片、文字、电子印章等），理论上文档库的结构细化到哪一级，通过 XML 数据库方式就可以访问到那一级的对象。

另外，在文档库系统中还存在一些描述文档库、文档集层次结构的元数据、权限数据和全局性数据、角色数据等。在不同的文档、或同一文档的不同页中可能会多次出现上述数据，依照上述存储方式进行文档数据存储时，这些数据会在包括它的不同数据片中被多次保存，这样，会降低存储空间的利用率。为提高存储空间的利用率，可以将这些数据进行单独保存，下面以保存全局性数据为例，说明一致性数据的保存方式：

步骤 1：提取文档库系统的全局性数据进行单独保存。

本步骤中，对全局性数据进行单独保存的方式可以包括：为全局性数据建立对应的层次结构；在建立的层次结构下，以 XML 文档、文件或二进制大对象的形式保存全局性数据。

步骤 2：在包括全局性数据的数据片中记录对全局性数据的引用。

应用上述方式进行全局性数据的保存后，即可以只对全局性数据保存一次，在其它数据片中不再记录单独保存的全局性数据，而是记录对该全局性数据的引用，从而大大提高了存储空间的利用率。对于其它诸如权限数据、元数据等一致性数据的保存方式和上述相同，这里就不再赘述。

在本实施例中，步骤 1 为全局性数据建立对应的层次结构可以为：将全局性数据对应 XML 数据库的树型结构中的某个层次，将全局性数据保存在该层次下。对于其它元数据、角色数据等也可以建立类似的与其对应的层次结构，并将相应的数据保存在该层次下。

在本发明第二本实施例中，通过文件系统方式表示文档库系统的树型结构。在该种方式下，本发明提供的文档库系统中文档数据的存储方法流程包括：

步骤 301：设置文档库系统的树型结构。

本步骤中，设置的文档库系统树型结构与实施例一中相同，这里就不再赘述。

表示文档库系统树型结构的方式可以为：利用文件系统的树型结构表示文档库系统的树型结构。

例如，对文档库系统设置文档集、文档、页、层和层内对象，即可以用文件系统的树型结构表示上述层次，如文档集中包括多个文档，文档中包括多个页，页中包括多个层，层中包括多个层内对象。可以利用下图表示文档库的树型结构：

- DosBase
  - DocList
    - Doc
      - Page
        - Object.sep

步骤 302：确定文档库系统中文档数据的存储粒度，并根据该存储粒度，将文档库系统中的文档数据拆成数据片。

本步骤中，确定文档数据的存储粒度、以及将文档数据拆成数据片的方式与实施例一中的相同，这里就不再赘述。

步骤 303：确定并保存拆成的数据片间的组织结构。

本步骤中，确定各个数据片间的组织结构，如，确定各个层数据所属的文档集、文档、页等。并将该组织结构以文件系统中目录的形式进行记录。

假定存储文档 Doc1 中的文档数据时，该文档 Doc1 位于 DocList1 下，Doc1 的文档结构为：包含 3 页，第一页包含 3 层，后两页包含一层。则在文档库中保存的组织结构如下所示：

- DosBase1
  - DocList1
    - Doc1

- Page1
  - Layer1.sep
  - Layer2.sep
  - Layer3.sep
- Page2
  - Layer1.sep
- Page3
  - Layer1.sep
- DocList1

步骤 304：保存拆成的数据片。

本步骤中，保存拆成的数据片的方式具体可以为：将拆成的各个数据片转换为 XML 文档保存到对应的目录下；或者，还可以将数据片作为一个文件系统中的文件或关系型数据库中的二进制大对象保存到对应的目录下。这里对应的目录可以是上述组织结构中相应的目录。

至此，本实施例提供的文档库系统中文档数据的存储方法流程结束。

在应用上述方法完成文档存储后，在进行文档访问时，可以通过指定文档集名+文档名+页号+层号+...的方式，快速定位到需要访问的存储粒度对象（如层对象），获取数据并返回给应用程序。

在本实施例中，也可以对文档库、文档集的元数据、角色数据、权限数据和全局性数据进行单独保存以提高存储空间的利用率，其具体方式与实施例一中步骤 1～2 类似。区别在于，对应本实施例中表示文档库系统树型结构的方式，为上述数据建立对应的层次结构的方式具体为：在文件系统的目录中建立一个特殊的目录作为全局性数据对应的层次结构，在该目录下保存全局性数据。具体地，上述全局性数据等可以是以 XML 文档、文件或二进制大对象的形式保存的；其它包括上述全局性数据的数据片不需要再单独保存，而只是包括对相应数据的引用，从而大大提高了存储空间的利用率。

当然，本实施例是以层为存储粒度为例进行说明的，事实上，还可以建

立更粗或更细的存储粒度对应的目录，以实现以更粗或更细的存储粒度对文档数据进行存储。

在本发明第三实施例中，通过关系型数据库方式表示文档库系统的树型结构。在该种方式下，本发明提供的文档库系统中文档数据的存储方法流程包括：

步骤 401：设置用于描述文档数据的树型结构。

本步骤中，设置的描述文档数据的树型结构与实施例一中相同，这里就不再赘述。

表示树型结构的方式为：利用关系型数据库中的树型结构表示文档库系统的树型结构。

例如，对文档库系统设置目录、文档、页三个层次，即可以关系型数据库中的 3 张表（目录表、文档表、页表）构建一个具有目录、文档、页三级结构的虚拟目录系统，其中，目录表的定义如表 1 所示，文档表的定义如表 2 所示，页表的定义如表 3 所示。

| 字段名 | 类型 | 说明 |
|---|---|---|
| DirID | INT | 目录的唯一 ID 号 |
| DirName | CHAR | 目录的名称 |
| ParentID | INT | 父目录的 ID（为 0 表示根目录） |

表 1

| 字段名 | 类型 | 说明 |
|---|---|---|
| DocID | INT | 文档的唯一 ID 号 |
| DocName | CHAR | 文档的名称 |
| DirID | INT | 文档所在目录的 ID |

表 2

| 字段名 | 类型 | 说明 |
|---|---|---|
| PageNo | INT | 页号（0 页代表全局性数据） |
| DocID | CHAR | 页所属文档的 ID |
| PageData | BLOB | 页数据 |

表 3

步骤 402: 确定文档库系统中文档数据的存储粒度，并根据该存储粒度，将文档库系统中的文档数据拆成数据片。

本实施例中，确定文档数据的存储粒度、以及将文档数据拆成数据片的方式与实施例一中的相同，这里就不再赘述。

步骤 403: 确定并保存拆成的数据片间的组织结构。

本步骤中，确定各个数据片间的组织结构，如，确定各个层数据所属的文档集、文档、页等。并将该层次结构以关系型数据库中相关表的形式进行记录。

假定存储文档 Doc1 中的文档数据时，该文档 Doc1 位于 DocList1 下，Doc1 包含 3 页，存储粒度为页。则在文档库中保存的组织结构和具体的数据片可以通过下面表 4～6 的相关表的方式存储。

| DirID | DirName | ParentID |
|---|---|---|
| 1 | DocBase | 0 |
| 2 | DocList1 | 1 |
| 3 | DocList2 | 1 |

表 4

| DocID | DocName | DirID |
|---|---|---|
| 1 | Doc1 | 2 |

表 5

| PageNo | DocID | PageData |
|---|---|---|
| 1 | 1 | Doc1 第 1 页数据 |
| 2 | 1 | Doc1 第 2 页数据 |
| 3 | 1 | Doc1 第 3 页数据 |

表 6

步骤 404: 保存拆成的数据片。

本步骤中，保存拆成的数据片的方式具体可以为：将拆成的各个数据片转换为 XML 文档保存到对应的相关表中；或者，还可以将数据片作为文件系统中的一个文件或关系型数据库中的二进制大对象保存到对应的相关表。

至此，本实施例提供的文档库系统中文档数据的存储方法流程结束。

在应用上述方法完成文档存储后，在进行文档访问时，可以通过指定文档集名+文档名+页号+层号+...的方式，快速定位到需要访问的存储粒度对

象（如页对象），获取数据并返回给应用程序。

在本实施例中，也可以对文档库、文档集的元数据、角色数据、权限数据和全局性数据等进行单独保存以提高存储空间的利用率，其具体方式与实施例一中步骤 1～2 类似。区别在于，对应本实施例中表示文档库系统树型结构的方式，为全局性数据等建立对应的层次结构的方式具体为：在关系型数据库的相关表中建立一个特殊的项作为相应数据对应的层次结构，在该目录下保存数据。

在具体实现时，对步骤 401 中设置的层次结构定义进行修改，将其中的存储粒度表定义（即页表定义）修改为表 7。

| 字段名 | 类型 | 说明 |
|---|---|---|
| PageNo | INT | 页号（0 页代表全局性数据） |
| DocID | CHAR | 页所属文档的 ID |
| PageData | BLOB | 页数据 |

表 7

在进行数据片存储后，形成的页表即变为表 8:

| PageNo | DocID | PageData |
|---|---|---|
| 0 | 1 | Doc1 全局性数据 |
| 1 | 1 | Doc1 第 1 页数据 |
| 2 | 1 | Doc1 第 2 页数据 |
| 3 | 1 | Doc1 第 3 页数据 |

表 8

可以看出，其中利用页号为 0 的页对应的相关表表示全局性数据对应的层次结构。可以将一致性数据作为关系型数据库中的二进制大对象、文件系统中的文件或数据库中的 XML 文档保存在对应的相关表中。其它包括一致性数据的数据片不需要再单独保存一致性数据，而只是包括对一致性数据的引用，从而大大提高了存储空间的利用率。

当然，本实施例是以页为存储粒度为例进行说明的，事实上，还可以建立更细的存储粒度对应的相关表，以实现以更细的存储粒度对文档进行存储。

通过上述本发明具体实施方式的描述，可以看出在本发明实施例的方法

中，首先设置文档库的树型结构；然后确定文档库系统中文档数据的存储粒度，并按照该存储粒度将文档库系统中的文档数据拆成数据片；确定并保存各数据片间的组织结构，并保存数据片。

图 2 为本发明实施例提供实现文档库系统中文档数据存储的装置结构示意图。如图 2 所示，本发明实施例的装置包括存储粒度确认单元 501、数据片拆分单元 502 以及记录单元 502。

其中存储粒度确认单元 501 确定文档库系统中文档数据的存储粒度。

数据片拆分单元 502 根据存储粒度确认单元 501 确定的存储粒度将文档数据拆成数据片。记录单元 502 记录数据片拆分单元 502 拆分的各数据片之间的组织结构，并在相应的组织结构下保存数据片拆分单元 502 拆分的各数据片。

应用本发明实施例提供的实现文档库系统中文档数据存储的方法及装置，可以将文档库系统中的文档数据拆分成更细的数据片并保存。这样，当用户检索或访问某文档时，可以快速、准确定位到最细存储粒度的级别，进行灵活高效地文档检索与内容抽取。对于基于网络的文档应用，可以只传输用户感兴趣的内容，降低了文档传输过程中的负担。

另外，由于本发明实施例进行文档数据存储时，可以将多种格式的文档进行统一保存，因此对于文档的互通互联带来了极大的便利。

以上仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

# 权利要求书

1、一种文档库系统中存储文档数据的方法，其特征在于，所述文档数据按树型结构描述，在存储所述文档数据时，包括:

将所述文档数据按所述树型结构拆成多个数据片;

以所述树型结构存储所述多个数据片，所述每一个数据片对应所述树型结构中的一个结点。

2、根据权利要求 1 所述的方法，其特征在于，将所述文档数据按所述树型结构拆成多个数据片为:

获取所述文档数据的存储粒度，将所述存储粒度作为最小拆分单位拆分所述文档数据为所述多个数据片。

3、根据权利要求 2 所述的方法，其特征在于，所述文档数据的存储粒度为所述树型结构中的一级。

4、根据权利要求 2 所述的方法，其特征在于，将所述文档数据按所述树型结构拆成多个数据片为:

将所述树型结构中所述存储粒度以上的每一结点的全局数据拆分为所述结点的子结点对应的一个或多个数据片。

5、根据权利要求 4 所述的方法，其特征在于，将所述每一结点的全局数据拆分为所述结点的子结点对应的一个或多个数据片为:

按照所述每一结点的全局数据的类别将所述全局数据拆分为所述结点的子结点对应的一个或多个数据片。

6、根据权利要求 1 所述的方法，其特征在于，利用可扩展标识语言 XML 数据库的树型结构表示所述树型结构;

以所述树型结构存储所述多个数据片为:

将所述树型结构的每个结点映射到 XML 数据库的树型结构中进行保存;

将所述多个数据片转化为数据库的 XML 文档，加入到所述 XML 数据库的树型结构中。

7、根据权利要求 1 所述的方法，其特征在于，利用文件系统的目录表示所述树型结构;

以所述树型结构存储所述多个数据片为:

将所述多个数据片作为文件系统的文件存储，并建立所述每个文件的目录。

8、根据权利要求 1 所述的方法，其特征在于，利用关系型数据库中的相关表表示所述树型结构;

以所述树型结构存储所述多个数据片为:

用所述关系型数据库中的相关表表示所述树型结构中每个结点与其父结点的信息;

将所述多个数据片作为所述关系型数据库的大对象存储。

9、根据权利要求 1 所述的方法，其特征在于，所述树型结构的根结点为文档级，所述叶子结点包括以下的一种或任意组合: 页、层、对象组和对象。

10、一种文档库系统中存储文档数据的装置，其特征在于，包括:

第一单元，用于将所述文档数据按所述树型结构拆成多个数据片;

第二单元，用于以所述树型结构存储所述多个数据片，所述第一单元拆分的每一个数据片对应所述树型结构中的一个结点。

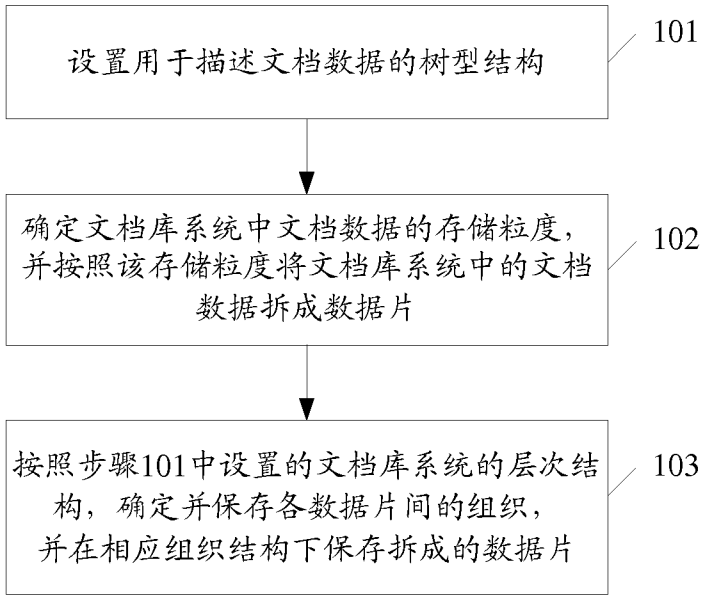

设置用于描述文档数据的树型结构
101
确定文档库系统中文档数据的存储粒度，并按照该存储粒度将文档库系统中的文档数据拆成数据片
102
按照步骤101中设置的文档库系统的层次结构，确定并保存各数据片间的组织，并在相应组织结构下保存拆成的数据片
103

图 1

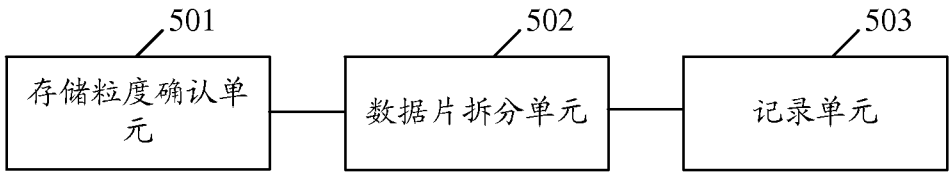

501
存储粒度确认单元
502
数据片拆分单元
503
记录单元

图 2

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2008/072010

CLASSIFICATION OF SUBJECT MATTER

G06F 17/30(2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: G06F 17/-; G06F 12/-; G06F 9/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNPAT, WPI, EPODOC, PAJ: document, file, data, split+, divide, division, partition, sub, child, manage, process, memory, store, node, tree

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| Y | CN 1244062 C (CANON KK) 01 Mar. 2006(01.03.2006) claims 1-11 | 1,9,10 |
| A | | 2-8 |
| Y | CN 1153162 C (FAST SEARCH & TRANSFER ASA) 09 Jun. 2004(09.06.2004) description page 2 line 10-page 5 line 28, figures 1-2 | 1,9,10 |
| A | | 2-8 |
| A | CN 1979472 A (MITSUBISHI ELECTRIC INFORMATION TECH) 13 Jun. 2007(13.06.2007) description page 9 line 1-page 10 line 4 | 1-10 |

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

- Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim (S) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 14 Nov. 2008(14.11.2008) | **27 Nov. 2008 (27.11.2008)** |
| Name and mailing address of the ISA/CN<br>The State Intellectual Property Office, the P.R.China<br>6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088<br>Facsimile No. 86-10-62019451 | Authorized officer<br>LIU,Li<br>Telephone No. (86-10)62411687 |

Form PCT/ISA/210 (second sheet) (April 2007)

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2008/072010

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| A | CN 1464440 A (TECHNOLOGY PROMOTING COMMITTEE) 31 Dec. 2003(31.12.2003) description page 2 line 4-page 4, figures 1-2 | 1-10 |
| P,X | CN 101021858 A (HUAWEI TECHNOLOGIES CO LTD) 22 Aug. 2007(22.08.2007) claims 1-12 | 1, 10 |

# INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2008/072010

| Patent Documents referred in the Report | Publication Date | Patent Family | Publication Date |
|---|---|---|---|
| CN 1244062 C | 01.03.2006 | WO 0229602 A1 | 11.04.2002 |
| | | AU 9351401 A | 15.04.2002 |
| | | EP 1323064 A1 | 02.07.2003 |
| | | KR 20030061819 A | 22.07.2003 |
| | | CN 1455901 A | 12.11.2003 |
| | | US 2004028049 A1 | 12.02.2004 |
| | | JP 2004510279 T | 02.04.2004 |
| | | AU 2001293514 B2 | 12.05.2005 |
| | | EP 1323064 B1 | 04.07.2007 |
| | | DE 60129232 E | 16.08.2007 |
| | | DE 60129232 T2 | 06.03.2008 |
| | | AT 366441 T | 15.07.2007 |
| CN 1153162 C | 09.06.2004 | NO 20002419 A | 13.11.2000 |
| | | AU 4321400 A | 21.11.2000 |
| | | WO 0068834 A1 | 16.11.2000 |
| | | BRPI 0010427 A | 19.02.2002 |
| | | CZ 20014002 A3 | 17.04.2002 |
| | | EP 1208465 A1 | 29.05.2002 |
| | | KR 20020006715 A | 24.01.2002 |
| | | CN 1360701 A | 24.07.2002 |
| | | NO 313347 B1 | 16.09.2002 |
| | | JP 2002544598 T | 24.12.2002 |
| | | AU 761169 B | 29.05.2003 |
| | | RU 2226713 C2 | 10.04.2004 |
| | | JP 3586429 B2 | 10.11.2004 |
| | | KR 100457830 B | 18.11.2004 |
| | | CA 2373453 C | 16.08.2005 |
| | | CN 1652108 A | 10.08.2005 |
| | | INCHENP 200101721 E | 07.09.2007 |
| | | US 7330857 B1 | 12.02.2008 |
| | | CN 100394424 C | 11.06.2008 |

# INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2008/072010

| Patent Documents referred in the Report | Publication Date | Patent Family | Publication Date |
|---|---|---|---|
| | | CA 2373453 AC | 16.11.2000 |
| CN 1979472 A | 13.06.2007 | WO 2007065355 A1 | 14.06.2007 |
| | | CN 1983251 A | 20.06.2007 |
| | | EP 1965309 A1 | 03.09.2008 |
| | | WO 2007065357 A | 14.06.2007 |
| | | WO 2007065356 A | 14.06.2007 |
| | | WO 2007065354 A | 14.06.2007 |
| | | WO 2007065353 A | 14.06.2007 |
| | | EP 1965310 A | 03.09.2008 |
| | | EP 1965327 A | 03.09.2008 |
| | | EP 1965308 A | 03.09.2008 |
| | | EP 1965314 A | 03.09.2008 |
| CN 1464440 A | 31.12.2003 | NONE | |
| CN 101021858 A | 22.08.2007 | WO 2008092341 A1 | 07.08.2008 |

# 国际检索报告

国际申请号
**PCT/CN2008/072010**

**A. 主题的分类**

G06F 17/30(2006.01) i

按照国际专利分类表(IPC)或者同时按照国家分类和 IPC 两种分类

**B. 检索领域**

检索的最低限度文献(标明分类系统和分类号)

IPC: G06F 17/-; G06F 12/-; G06F 9/-

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)CNKI, CNPAT, WPI, EPODOC, PAJ:文件, 文档, 数据, 分割, 划分, 拆分, 拆成, 子, 级, 层, 段, 块, 管理, 处理, 存储, 储存, 节点, 结点, 树型, 树形, document, file, data, split+, divide, division, partition, sub, child, manage, process, memory, store, node, tree

**C. 相关文件**

| 类　型* | 引用文件，必要时，指明相关段落 | 相关的权利要求 |
|---|---|---|
| Y | CN 1244062 C (佳能柱式会社) 01.3 月 2006(01.03.2006) 权利要求 1-11 | 1,9,10 |
| A | | 2-8 |
| Y | CN 1153162 C (快速检索及传递公司) 09.6 月 2004(09.06.2004) 说明书第 2 页第 10 行-第 5 页第 28 行，附图 1-2 | 1,9,10 |
| A | | 2-8 |
| A | CN 1979472 A (北京书生国际信息技术有限公司) 13.6 月 2007(13.06.2007) 说明书第 9 页第 1 行-第 10 页第 4 行 | 1-10 |
| A | CN 1464440 A (财团法人资讯工业策进会) 31.12 月 2003(31.12.2003) 说明书第 2 页第 4 行-第 4 页，附图 1-2 | 1-10 |
| P,X | CN 101021858 A (华为技术有限公司) 22.8 月 2007(22.08.2007) 权利要求 1-12 | 1, 10 |

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

| 国际检索实际完成的日期 | 国际检索报告邮寄日期 |
|---|---|
| 14.11 月 2008(14.11.2008) | **27.11 月 2008 (27.11.2008)** |
| 中华人民共和国国家知识产权局(ISA/CN)<br>中国北京市海淀区蓟门桥西土城路 6 号 100088<br>传真号：　(86-10)62019451 | 受权官员<br>刘力<br>电话号码：(86-10) **62411687** |

**国际检索报告**
关于同族专利的信息

国际申请号
**PCT/CN2008/072010**

| 检索报告中引用的专利文件 | 公布日期 | 同族专利 | 公布日期 |
|---|---|---|---|
| CN 1244062 C | 01.03.2006 | WO 0229602 A1 | 11.04.2002 |
| | | AU 9351401 A | 15.04.2002 |
| | | EP 1323064 A1 | 02.07.2003 |
| | | KR 20030061819 A | 22.07.2003 |
| | | CN 1455901 A | 12.11.2003 |
| | | US 2004028049 A1 | 12.02.2004 |
| | | JP 2004510279 T | 02.04.2004 |
| | | AU 2001293514 B2 | 12.05.2005 |
| | | EP 1323064 B1 | 04.07.2007 |
| | | DE 60129232 E | 16.08.2007 |
| | | DE 60129232 T2 | 06.03.2008 |
| | | AT 366441 T | 15.07.2007 |
| CN 1153162 C | 09.06.2004 | NO 20002419 A | 13.11.2000 |
| | | AU 4321400 A | 21.11.2000 |
| | | WO 0068834 A1 | 16.11.2000 |
| | | BRPI 0010427 A | 19.02.2002 |
| | | CZ 20014002 A3 | 17.04.2002 |
| | | EP 1208465 A1 | 29.05.2002 |
| | | KR 20020006715 A | 24.01.2002 |
| | | CN 1360701 A | 24.07.2002 |
| | | NO 313347 B1 | 16.09.2002 |
| | | JP 2002544598 T | 24.12.2002 |
| | | AU 761169 B | 29.05.2003 |
| | | RU 2226713 C2 | 10.04.2004 |
| | | JP 3586429 B2 | 10.11.2004 |
| | | KR 100457830 B | 18.11.2004 |
| | | CA 2373453 C | 16.08.2005 |
| | | CN 1652108 A | 10.08.2005 |
| | | INCHENP 200101721 E | 07.09.2007 |
| | | US 7330857 B1 | 12.02.2008 |
| | | CN 100394424 C | 11.06.2008 |
| | | CA 2373453 AC | 16.11.2000 |
| CN 1979472 A | 13.06.2007 | WO 2007065355 A1 | 14.06.2007 |
| | | CN 1983251 A | 20.06.2007 |
| | | EP 1965309 A1 | 03.09.2008 |
| | | WO 2007065357 A | 14.06.2007 |
| | | WO 2007065356 A | 14.06.2007 |
| | | WO 2007065354 A | 14.06.2007 |
| | | WO 2007065353 A | 14.06.2007 |
| | | EP 1965310 A | 03.09.2008 |
| | | EP 1965327 A | 03.09.2008 |
| | | EP 1965308 A | 03.09.2008 |
| | | EP 1965314 A | 03.09.2008 |
| CN 1464440 A | 31.12.2003 | 无 | |
| CN 101021858 A | 22.08.2007 | WO 2008092341 A1 | 07.08.2008 |