植物遗传资源学报 2011, 12(1): 154-157
Journal of Plant Genetic Resources

基于属性分离存储的种质数据组织方法

唐 鹏, 方 沩, 司海平, 曹永生
(中国农业科学院作物科学研究所, 北京 100081)

摘要: 传统的作物种质数据组织方法, 针对不同作物种类建立不同数据表, 这种方法已不能有效适应种质数据综合分析的需要。本文提出了一种基于属性分离存储的种质数据组织方法, 根据种质的每个属性分别建立数据表, 各属性间没有从属关系。该方法可统一数据查询操作, 优化查询过程, 提高分析效率, 具有灵活、可扩展的特点, 可以方便地集成与种质分析相关的数据, 适用于种质资源分布式数据库和相关信息系统的建立。

关键词: 属性分离存储; 种质数据; 数据组织

A Germplasm Data Organization Method Based on Property-Separation Storage

TANG Peng, FANG Wei, SI Hai-ping, CAO Yong-sheng
(*Institute of Crop Sciences, Chinese Academy of Agricultural sciences, Beijing 100081*)

Abstract: Traditional germplasm data organization methods, in which different tables concerned different crops, can not meet the need for germplasm synthetic data analysis. This paper raises a new organization method based on property-separation storage, in which all property data were put in the same level without affiliation. With this method, the data query procedures can be unified with high efficiency. This method is flexible and extendable, which can be applied in germplasm-related information system.

Key words: Property-separation storage; Germplasm data; Data organization

有效的种质数据组织是建立种质数据管理系统、高效地进行数据分析的前提[1]。始建于 1986年的国家农作物种质资源数据库, 应用了基于作物种类的种质数据组织方法, 按作物种类建立了不同的数据表。这种数据组织方法简便、直观, 但仅适用于单作物 (例如水稻) 数据的研究, 不适合多作物数据的综合分析。对于多作物数据的分析需要进行大量的数据有效性验证、统一性验证、数据表联合等操作, 其中许多操作必须通过数据库管理系统外部编程实现[2]。例如从不同数据表中选择高蛋白质含量的种质, 需要执行很多跨数据表操作。随着作物种类、数量以及对各种数据综合分析需求的增加, 原有的数据组织方法已不能适应分析的需要。为此, 本文结合传统种质数据组织方法的优点, 提出了基于属性分离存储的种质数据组织方法。

1 数据组织方法

基于属性分离存储的种质数据组织方法, 根据种质的每个属性[3-4]分别建立数据表[5], 各属性间没有从属关系, 每种作物的不同属性存储于不同的数据表, 不同作物的相同属性存储于同一数据表中。例如, 水稻的原产地、蛋白质含量、稻瘟病等属性[6]分别存储在不同的数据表中; 小麦的原产地、蛋白质含量、条锈病等属性也分别存储在不同的数据表中。小麦的原产地、氨基酸含量等属性与水稻的相同属性保存在同一表中。

收稿日期: 2010-03-20 修回日期: 2010-10-22
基金项目: 国家科技基础性工作专项 (2006FY110700)
作者简介: 唐鹏, 在读硕士, 研究方向为管理信息系统, 计算计网络。E-mail: tangpeng@caas.net.cn
通讯作者: 曹永生, 研究员, 主要从事种质资源信息系统、图象识别、GIS 领域研究。E-mail: caoys@mail.caas.net.cn

每个属性表中的属性数据都需要有一个主键。主键的不同值表示不同的种质,种质数据的不同属性通过主键关联[7-9]。

这种方法的实现需要两个步骤: 数据表建立和数据存储。

1.1 数据表建立

该方法中的数据表分为 3类: 属性表、种类表、辅助表。每类数据表包含的具体数据表如表 1所示。

表 1 每类数据表包含的具体数据表

Table 1 Details of each type of table

类型 Table type	表名 Table name	作用 Effect
属性表	统一编号表 (tb_accnum)	保存每份种质的编号
	作物名称表 (tb_croptype)	保存每份种质所属的作物名称
	属性表 (tb_propertyX)	保存每个属性的数据
种类表	tb_datatype	保存所有作物的属性表名称
辅助表	tb_depict	保存每个属性表描述

属性表用来保存属性数据。每个属性表的表名不同,但表结构均只包含两个字段,其中一个字段为标识 (acc_id), 另一个字段 (property) 存储属性数据。

在属性表中,有两个表具有特殊意义。统一编号表 (tb_accnum) 记录每一份种质的统一编号,并在数据插入时自动产生一个唯一的标识 (acc_id)。数据分类表 (tb_croptype) 保存每份种质所属的作物名称。

种类表 (tb_datatype) 保存所有作物的种类 (acc_type) 及其属性表名称 (property_table_name)。

辅助表 (tb_depict) 保存属性表名称 (property_table_name) 及其属性表描述内容 (property_table_depict)。

1.2 数据存储

种质数据存储要经过 4个步骤: 新建作物种类、生成标识、插入作物种类属性、插入其他数据。

1.2.1 新建作物种类 在初始状态下,种类表 (tb_datatype) 和属性表均为空表。存储数据时,需新建一个作物种类。新建作物种类时只需在 tb_datatype表中插入作物名称及其所有属性表名,每个属性表名为一条记录。具体内容如表 2所示。

表 2 tb_datatype表示例

Table 2 Sample data in tb_datatype

acc_type	property_table_name
小麦	tb_protein
小麦	tb_abc
小麦	tb_dd
…	…
水稻	tb_dd
水稻	tb_protein
…	…

1.2.2 生成标识 在统一编号表 (tb_accnum) 表中插入一条记录,记录种质的统一编号和自动生成的标识 (acc_id)。

1.2.3 插入作物种类属性 在种类表 (tb_croptype) 中插入一条记录,记录该数据的分类和 acc_id。

1.2.4 插入其他数据 读取 tb_datatype表,获取该分类的属性表名称。若无该属性表,则新建此属性表并将描述信息插入 tb_datatype表。重复获取数据的每个字段内容,将该数据的 acc_id与属性数据插入对应属性表。

数据的修改和删除只需知道数据的标识和待修改的属性。根据标识对相应的属性表进行 update或者 delete操作。修改种类种质数据的结构时,只需直接在 tb_datatype中修改数据,无需修改表结构。

在实际插入、修改、删除数据的过程中,可应用现代关系数据库中事务 (Transactions) 特性存储数据以保证数据的完整性。

2 方法应用

2.1 分作物查询

分作物查询时要对种类表 (tb_datatype) 表执行查询,查询条件为 acc_type= 作物种类,可获得待查询作物的所有属性表表名,根据获取的表名和统一编号表 (tb_accnum) 产生查询。其结果为单作物种类的所有数据。其中, tb_accnum为主表,查询结果中包含了 tb_accnum中特定种类的所有数据及所有属性表中与主表左联 (left out join) 的数据。在查询时,若不指定其他查询条件,则显示待查作物的所有属性数据,也可任意指定某属性不参与查询。若指定其他作物种类数据的属性参与查询,则查询结果为 NULL。操作时数据表关联如图 1所示。

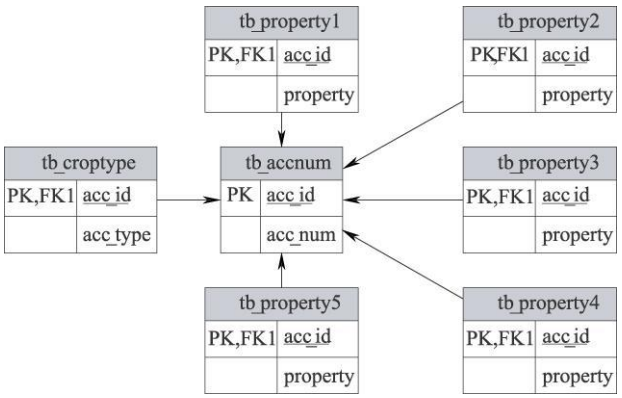

图 1 单作物类型数据获取

Fig 1 Data query for a single crop type

tb_accnum 为主表, 其余表为从表。从表通过查询 tb_datatype 获取

程序设计时显示数据的“字段名”以及查询语句生成时指定别名 (as)操作, 需要一个单独的查询获取属性表的描述信息, 其数据关联如图 2所示。

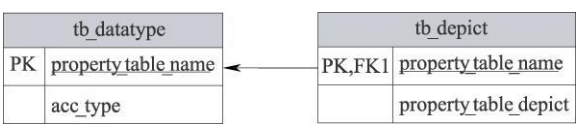

图 2 属性表描述信息获取

Fig 2 Accquire description for property tables

tb_datatype为主表, tb_depict为从表

2 2 分属性查询

针对单属性的多作物数据获取, 只需按图 3所示数据关联条件生成查询。

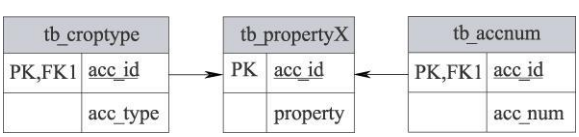

图 3 单属性数据获取

Fig 3 Data query for a single property

tb_propertyX 表中的 property为待查询属性。 tb_propertyX为主表, tb_accnum和 tb_croptype为两个从表

若要包含多个属性, 可联接 (join) 多个表。若查询中其他属性表与主属性表的关系为内联 (inner join), 则查询结果为所有属性表均不为空的数据; 若查询中其他属性表与主属性表的关系为左联 (left join), 则查询结果包含主表中所有匹配数据, 子表中的不匹配数据将被 NULL填充。操作时的数据表关联如图 4所示。

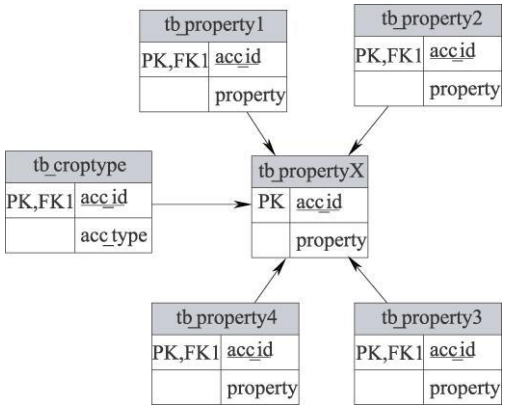

图 4 多属性数据获取

Fig 4 Data query for multiple properties

tb_propertyX 表中的 property为待查询属性。 tb_propertyX为主表, 其余表为从表

2 3 数据综合分析

对数据的各种分析首先是对数据的各种查询操作。由于采用了属性分离存储的数据存储方式, 种质数据的属性表都是平等的关系, 不存在从属关系, 所以这种存储方法非常适合对不同种类种质数据的综合分析。表 3列举了一些实际分析时产生的一些查询操作。

表 3 具体应用举例

Table 3 Examples of specific data queries

应用实例 Example	执行操作 Corresponding operation
筛选每个省中蛋白质含量最丰富的一份种质材料	select max(b. property) as las, a. acc_num, c. property province from tb_accnum a left join tb_protein b on a. acc_id = b. acc_id left join tb_province c on c. acc_id = a. acc_id group by a. property
筛选高海拔地区种质材料 (海拔 3000m 以上) 以供分析	select a. * , b. property latitude from tb_accnum a left join tb_hb b on a. acc_id = b. acc_id where b. property > 3000
每个省的种质份数	select count(*), tb_province. property province from tb_province group by property
某一材料采集地点的其他种质材料	select b. * from tb_accnum a left join tb_location b on a. acc_id = b. acc_id join tb_accnum c on c. acc_num = ‘xxx’ and d. property = b. property join tb_location d on d. acc_id = c. acc_id

通过举例可以发现, 查询语句随着不同的查询需求, 长度、查询条件有所变化, 但是其基本格式保持统一。对于常用的查询操作, 可编写存储过程批量化的实现。

表 4对比了传统的种质数据库组织方法与本文

所述方法的特点。

表 4 传统种质数据库组织方法与基于属性分离存储的种质数据组织方法比较

Table 4 Analysis between the traditional and property-separation-storage-based germplasm data organization methods

对比项目 Features	传统方法 Traditional method	基于属性分离存储的方法 Property-separation-storage-based method
数据表数量	少	多
数据表结构	复杂、结构各异	简单、统一
联合分析	需外部编程支持[1]	数据库管理系统内部运算实现
扩展	每种新作物都需要新建数据表[2]	属性不存在时新建数据表, 新作物无需新建数据表
可迁移性	适用于各类简单/关系数据库管理系统[1-2]	适用于关系数据库管理系统
数据查询时间损耗	单表查询速度快, 多表查询速度慢, 外部程序调用数据库需大量时间	单表/多表查询时间基本一致, 没有外部程序调用损耗

3 讨论

本文充分利用了现代关系数据库的特性, 提出了基于属性分离存储的种质数据组织方法。该方法在保留了传统种质数据组织方法优点[1-2]的基础上, 可统一数据查询操作, 优化查询过程, 提高分析效率, 具有灵活、可扩展的特点, 能将绝大多数对数据的操作在现在流行的数据库管理系统(例如 Oracle, SQL Server, MySQL, PostgreSQL等)中完成, 可以方便地集成与种质分析相关的数据, 适用于种质资源相关信息系统的建立。

该方法采用了类似数据库纵向分片(DB vertical partition)技术, 可以极大提高数据库的数据容量和检索效率, 并能方便地实现基于数据库分片技术的分布式数据库(Distributed database)[10]。

该方法已应用于云南及周边地区农业生物资源调查信息系统(项目编号: 2006FY110700)、中国沿海地区抗逆农作物种质资源调查信息系统和中国作物种质资源信息系统(CGRIS), 其灵活性、可扩展性、稳定性得到了检验。该方法不仅适用于作物种质数据的组织, 还适用于林木、药用植物、动物和微生物等种质信息的组织。

参考文献

[1] 曹永生, 陈育, 孔繁胜. 中国作物种质资源信息共享网络的建立[J]. 资源科学, 2001(1): 46-48

[2] 张贤珍, 孔繁胜, 曹永生. 农业数据库系统程序设计[M]. 北京: 农业出版社, 1991

[3] 张贤珍, 杨克钦, 荣绛平. 农作物品种资源信息处理规范[M]. 北京: 农业出版社, 1990

[4] 董玉琛, 刘旭. 农作物种质资源技术规范丛书[M]. 北京: 中国农业出版社, 2007

[5] 陈启才. 查询重写关系数据库查询优化技术[J]. 电脑编程技巧与维护, 2009(4): 40-42

[6] 曹永生, 张贤珍. 国家种质资源数据库的优异稻种资源信息[J]. 作物品种资源, 1993(3): 19-21

[7] 胡天磊. 修炼SQL[J]. 程序员, 2007(3): 52-53

[8] 任永昌, 邢涛, 陈晓纪. 基于关系数据库的模型库系统研究[J]. 渤海大学学报(自然科学版), 2008, 29(2): 180-184

[9] 王丽华. 关系数据库范式及应用[J]. 内江科技, 2008, 29(7): 117-118

[10] 杨晶, 刘天时, 马刚. 分布式数据库数据分片与分配[J]. 现代电子技术, 2006, 29(18): 119-121

(上接第 153页)

参考文献

[18] 刘勇, 林刚, 何光源, 等. 两个油菜种的染色体核型分析[J]. 华中科技大学学报(自然科学版), 2005, 33(3): 119-121

[19] 韩清芳, 贾志宽. 紫花苜蓿种质资源评价与筛选[M]. 杨凌: 西北农林科技大学出版社, 2004

[20] 张晓红. 南苜蓿和紫花苜蓿染色体的核型分析[J]. 安徽农学通报, 2009, 15(7): 51-52

[21] 张凤仙, 毕玉芬, 王晓云, 等. 云南野生苜蓿与引进苜蓿的核型分析[J]. 云南农业大学学报, 2008, 23(7): 431-435

[22] 贾纳提维纳汗, 李保军, 郑晓红. 四种豆科牧草的染色体核型分析[J]. 草业科学, 1996, 13(4): 11-13

[23] 聂谷华, 廖亮, 方亮, 等. 国产毛茛属植物五种一变种的核型[J]. 云南植物研究, 2007, 29(3): 316-322

[24] 陈家宽, 杨继. 植物进化生物学[M]. 武汉: 武汉大学, 1994

[25] 张为民. 四种紫花苜蓿的核型分析[J]. 山西农业大学学报, 2006, 26(1): 73-76

[26] 孙爱群, 向红, 田应洲, 等. 珍珠菜属四种植物的核型分析[J]. 武汉植物学研究, 2007, 25(5): 509-512