一种基于目录和关系的混合数据模型

赵 曦[1] 陈建阳[2]
(上海金融学院信息管理系 上海 200135)[1] (同济大学 上海 200127)[2]

摘 要 目录和关系数据库分别在架构、安全、事务处理等方面各有优点,而许多信息管理系统,特别是互联网应用(电子商务、分布应用)系统中,既需要目录数据结构的与组织资源架构的一致性和分布特征,也需要关系数据对事务处理、数据挖掘的优势,两种数据结构的结合可以使系统数据更加结构化、应用开发更加高效、变更更加灵活。本文提出了一个基于 LDAP/X500 目录和 SQL 数据操作语言的关系数据库的混合数据模型框架,来解决一些具有组织结构、资源管理,又有事务处理、数据分析应用系统的数据框架问题。

关键词 目录,LDAP/X500,关系数据库,SQL 操作语言

A Hybrid Data Framework of Relational and Directory Models

ZHAO Xi CHEN Jian-Yang
(Department of Information Management, Shanghai Finance University, Shanghai 200135) (Tongji University, Shanghai 200127)

Abstract Directory and relational data base have both advantages in architecture, security, transaction processing etc, but some information management systems, especially online applications based on internet such as e-commerce, distributed applications need the uniform and distribution proprieties provided by directory as well as business transaction processing, data analysis by relational data bases. This article presents a hybrid data framework based on both LDAP/X500 directory and Sql related relational dada models, this framework can help in applications which require organization, resource management and business transaction data processing.

Keywords Directory, LDAP/X500, Relational data base, SQL language

1 引言

LDAP 目录数据(directory)和 RDB 关系型数据标准的技术与数据库产品广泛应用于各行业的信息管理系统中。随着互联网应用模式的发展和业务数据处理的复杂程度增加,以及业务系统的服务交叉和大量数据交换对数据处理能力、服务稳定性和系统安全性提出了更高的要求,同时信息管理规划需要对现有管理系统进行业务和数据方面的统一,而这些系统在数据架构、开发技术和业务主体方面的差异使得对它们的整合难度较大,使用面向服务架构 SOA 无异于重新开发系统,使用 WebService 接口能够完成部分业务协同和数据共享,但缺乏统一的操作规范和全面的解决方案。本项目致力于构建一个以 LDAP 目录和 RBD 关系为基础的综合数据模型的管理系统数据 DRDB 架构,并提出一个混合数据操作脚本定义。它能够更加完善地描述网络化、分布式的大型信息管理系统的数据结构,提高数据处理和数据迁移效率,能够统一地描述和处理组织架构、资源管理、业务处理和数据分布的信息管理系统数据模型。

2 目录和关系数据库

以 LDAP/X500 标准为基础的目录数据结构与实际应用中的管理架构(行政管理、企业经营、资源组织等)完全吻合而得到广泛应用,特别在大型分布式管理、系统权限和用户管理方面。目录数据管理系统可以灵活构建跨区域、逐步完善的树状数据网络,通过分布、复制机制实现数据的统一和共享。

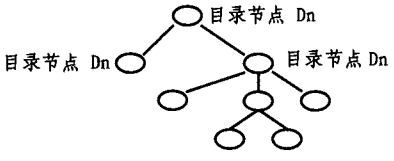

图 1 目录数据模型示意图

目录数据库具有分布性强、数据查询效率高、信息安全机制完备、数据迁移灵活和平台无关性强等特点。目录数据能够很好地处理"数据架构"、"数据权威"和"数据定位"问题,但其数据处理功能较弱,在数据类型、批量处理、事务处理、数据挖掘方面的能力不足。

以关系模型和 SQL 数据操作语言为基础的关系数据库广泛应用于各行业的信息管理系统,成为主流的数据库服务技术。关系数据库通过数据表格和表格之间的连接关系来描述信息管理系统的业务数据格式,通过 SQL 语言来操作数据,关系数据库产品和 SQL 语言提供了强大的数据定义、数据存储、数据检索和事务处理功能,同时提供了存储过程、事件触发和数据挖掘功能。

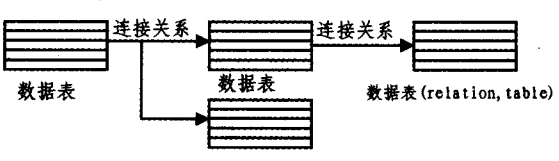

图 2 关系数据模型示意图

与目录型数据相对应,关系型数据库对业务数据、事务管

理、批量处理、统计分析的处理能力强、效率高，但对跨区域数据整合、与管理架构保持统一、数据迁移，以及访问安全控制方面有一定的先天不足，或通过对问题的复杂化处理来满足数据服务的需求。

目前，信息管理系统正朝着网络化、分布化、大容量、多事务、个性化方向发展，如大型企业集团的 ERP、物流、行政管理、网上银行、电子商务等应用，需要有一种能够兼备目录和关系数据库优点的数据模型和操作，来满足日益变化的信息管理要求。

3 目录和关系数据混合模型

本文提出一个综合目录和关系数据模型的数据信息管理架构(Directory / Relational Data Base Frame)的设想，以及一个操作操作该数据架构的脚本语言定义，并通过开发脚本语言解释程序来验证提出的混合数据架构的合理性和可行性。LDAP/X500 目录通过节点表识(Dn, Distinguished Name)、类型(ObjectClass)、数据分布(distribution)、数据复制(replication)，以及数据权限机制来描述和实现企业、单位组织架构和资源分布，通过基于 SQL 操作语言的关系型数据库(RDB)来实现业务数据处理和事务管理，从而能够提高和改进系统分析的质量。

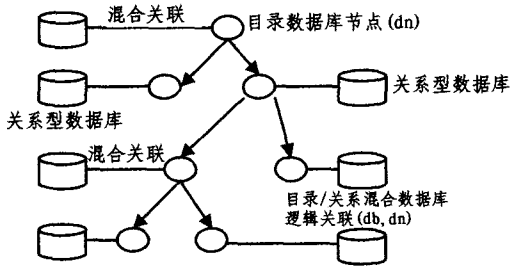

图 3 目录/关系混合数据模型示意图

混合关联的实现要求关系数据库记录对应目录节点的 Dn(唯一的)，可以通过关系数据库参数或表格字段来实现。以混合数据框架为基础的信息管理系统将在以下几个方面得到改进：

1)通过目录数据结构保障基本信息数据的权威性和唯一性，提高数据结构与业务组织、信息处理模型逻辑的吻合程度；

2)提高信息管理系统数据服务功能的效益和性价比；现有的大部分数据管理应用是使用关系数据库来实现，而相当多的信息管理和业务处理应用并不需要大量的数据事务处理和数据挖掘功能；

3)提高数据服务系统对业务组织变化，业务流程变化和网络结构变化的适应能力，组织架构和业务流程的变化都会引起信息管理系统的修改和升级。相当数量的系统需要重新设计才能满足要求，造成不必要的重复投资和资源浪费。

3.1 操作脚本定义

与现有模型 LDAP 和数据操作语言(SQL)的结合，协议规范的 API 与数据服务的结合，目录/关系混合的连接信息和标识，分布数据和分布事务处理。

1)混合数据查询

On dn LdapType LdapFilter do {SQL statement};

其中：On—混合数据操作关键字

Do—进行关系数据操作关键字

Dn—数据操作的目录节点 DN

LdapType—目录节点操作类型，分别取以下值：

BaseEntry：数据操作只针对目录节点 Dn 进行

OneLevel：数据操作对 Dn 的下一级目录节点进行

subEntry：数据操作对 Dn 的下级目录节点进行

LdapFilter—对目录节点 Dn 进行 LdapType 操作的限制过滤条件

Sqlstatement—符合 SQL [2000]标准的关系数据操作语句

下面是一个混合数据查询脚本例子：

```
On ou=Shanghai,ou=cn SubEntry objectclass=subcompany do
{ select a. client_id,a. client_credit from client a
  order by a. client_credit desc
}
```

上述脚本含义为：对 Dn 为 ou=Shanghai,ou=cn 节点的所有下属节点(子公司)，分别查询表格 client 的信息。数据库参数和连接(池)由脚本执行 API 函数处理。

2)数据结构调整(移动、合并)

Move dbx from Dn1 to Dn2

将数据库 dbx 从关联节点 Dn1 移动到 Dn2，并修改 Dn1，Dn2 的连接属性在不涉及结构和分布的情况下，关系数据库的操作使用常规的 SQL 操作语言，目录数据库使用 LDAP 规范接口和 API。由应用系统建立和维护关系数据库(rdb)的连接和连接池管理，并将数据库状态(可用、断开 等)通告对应的关联目录节点。

3.2 混合数据 Cache 和指针管理

混合数据查询操作会在 cache 内获得一个数据视图(图 4)。

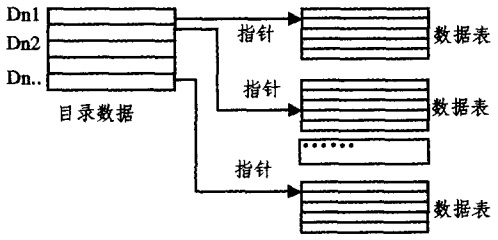

图 4 目录/关系混合数据操作视图

在对数据表进行处理时(翻页，跳转等)，需要操作多个指针和保存它们的状态。查询结果记录总数为各个 Dn 对应数据表记录的和。翻页或跳转时按顺序移动，到底后再移动下一个 Dn 对应的数据表指针。

混合数据模型的分布特性涉及分布事务处理的问题，脚本执行 API 函数将在应用(application)中进行确认(commit)和回滚。

4 混合数据模型应用和测试

我们使用上述混合数据框架原理应用到多个进行树状组织架构管理和业务数据处理的信息系统，涉及公安、农业、城市管理等领域。其中一个实际应用-建设项目招投标信息管理系统，用于管理城市建设管理机构(市/区县)和项目招投标的信息化管理，需要保持架构的统一和分布式的业务处理，其数据架构如图 5 所示。

目录数据库采用 OpenLdap2.0，关系数据库采用 SqlServer2000，应用开发采用 B/S 模式和 DCS 软件开发中间件。

(下转第 149 页)

从上表可以看出，尽管一些关键词的 tf 值(训练文本 k 中词语 i 出现的次数)和 df 值(训练集中出现词语 i 的文本数)都很低，如“计划生育”、“党风”，这些词在 Okapi 权重函数中不会取得较高的分值，但它们不为 0 的 $delt$ 值可以给它们一倍左右的权值提升。同时对于“河南省”等本身权值较高的词汇，其 $delt$ 值为 0，提升滤波器对它们没有影响。

2.5 词汇权重离散化

在离散化之前，先将权重小于等于 0 的项(干扰项)滤去。这里采用简单的等距离离散化方法，将权值离散化为 0 到 10 共 11 个级别。

此时上表的结果如下：

表 2 新闻文本的 tf、df、delt 及离散化后权重值计算结果示例表

关键词	tf 值	Df 值	Delt 值	离散化后的权重
新蔡县	3	5	1.3	5
河南省	4	10	0	2
计划生育	1	4	2.4	5
人口	4	12	0	1
党风	1	4	2.4	7
建设	1	12	0	0
……				

3 利用粗糙集理论获取规则

粗糙集理论作为一种新的机器学习方法[11]，已经在很多方面进行了应用。其中文本分类是粗糙集的几个经典应用领域之一。我们采用粗糙集作为文本分类工具。在本次实验中，选取 100 篇国际新闻，100 篇国内新闻。各留 100 篇做测试。在前期分词、权重过滤之后，尚有 3216 个特征词汇，我们选取权重排名前 1000 个特征词汇，在此基础上进行属性约简，得到分类规则。

在 ROSETTA 粗糙集平台[12]上进行知识约简获取规则，在 200 篇测试集上做开放测试，结果如下：

表 2 新闻文本的 tf、df、delt 及离散化后权重值计算结果示例表

参数选取	S=3，a=0.03		S=0，a=0(传统方法)	
查全率	国际新闻	91.9%	国际新闻	87.2%
	国内新闻	94.3%	国内新闻	91.6%

由上表可见，改进后算法提高了开放测试的查全率。

总结和展望 通过滤波器器的方法将隐藏未知子类的特征词汇赋予较高的权值，使我们能够充分利用文本分类中类别的结构信息，提高查全率。在我们提升小规模子类特征的时候，虽然确实能把小规模子类的一些重要特征提升起来，但同时也提升了一些并不重要的特征，使噪声加大，因而如何选择合适的参数使结果最佳将是我们进一步的研究工作。

参 考 文 献

1 Sabstiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34:1～47
2 Lewis D R. A comparison of two learning algorithms for text categorization [A]. In: Symposium on Document Analysis and IR,[C]. Las Vegas: University of Nevada Press, 1994. 81～93
3 Yang Y, LIU X. A re-examination of text categorization methods [A]. In: Proc. of the 22nd Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval[C]. Berkeley: ACM Press, 1999. 42～49
4 杨丽华，戴齐. KNN 文本分类算法研究. 微计算机信息，2006，22
5 Mccallum A, Nigam K. A comparison of event models for naive Bayes text classification [A]. AAAI-98 Workshop on Learning for Text Categorization [C], Madison: AAAI Press, 1998. 22～28
6 Nigam K, Lafferty J, Mccalum A. Using maximum entropy for text classification [A]. IJCAI-99 Workshop on Machine Learning for Information Filtering [C], Stockholm: IJCAI Press, 1999. 35～42
7 卢娇丽，郑家恒. 基于粗糙集的文本分类方法研究. 中文信息学报，2005，19(2)
8 李钝，梁吉业. 利用聚类和粗糙集进行文本分类研究. 计算机工程与应用，2003，07-0186-03
9 胡荣，罗庆云. kNN 算法在文本分类中的改进. 南华大学学报(自然科学版)，2005，19(3)
10 Robers S E, Walker S. Okapi/Keenhow at TREC8 [A]. In: E. M. Yoorhees, D. K. Harmann, eds. Proceedinging of Eighth Text Retrieval Conference(TREC-8)[C], Gaithersburg, 2000
11 苗夺谦. Rough Set 理论及其在机器学习中的应用研究[D]. 北京：中国科学院自动化研究所，1997
12 http:// rosetta. lcb. uu. se/

(上接第 125 页)

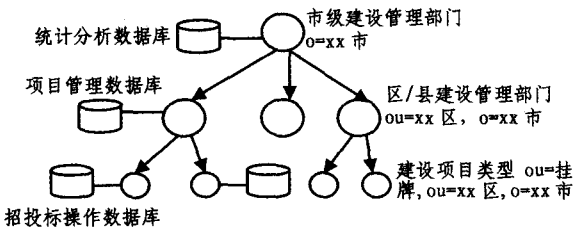

图 5 建设项目招投标管理数据框架

目录数据分布通过网络将市级管理部门与区/县管理部门连接在一起，区县部门通过招投标管理系统进行业务操作和关系数据库更新，包括项目公共、资格审查、评分规则、评标计算和结果公示。不同类型项目(招标、挂牌、拍卖)的评分规则和标准、专家分组和数量是不同的。

通过应用系统的功能开发，在混合数据框架下的每个目录节点(市/区/项目类型)，管理部门能够进行逐级管理和汇总分析，体现了本文提出的混合数据框架的可行性和优势。

结论 由于业务复杂程度的加大，以及各种专业化业务处理系统的增多，企业和单位进行信息系统规划和资源整合的难度很大。本文提出一个利用目录和关系数据结构优势的混合数据框架，可以既保持业务处理系统灵活性和独立性，又能够进行组织架构及其变化的管理，使数据做到有效的统一、分布和共享，对信息系统建设和整合中的数据结构化程度，数据组织和处理效率有一定的特色和优势。

参 考 文 献

1 王芳，张顺达，等. 基于 LDAP 的对象存储系统元数据的组织与管理. 计算机工程与科学，2007(3)
2 薛宏智，王俊，等. 基于 LDAP 的企业访问控制系统设计与实现. 计算机与信息技术，2006(Z1)
3 李明，连乔，等. SQL 标准符合性测试的框架. 计算机工程与应用，2003(20)
4 王建芳，阎保平，等. 目录的数据管理模型的研究与实现. 计算机工程，2007(10)
5 罗艳兵，蔡鸿明. 应用企业数据模型解决信息孤岛的研究. 新西部(下半月) 2007(1)
6 RFC2253. Lightweight Directory Access Protocol (v3)
7 ISO/IEC 9075:1992, Database Language SQL