第37卷 第19期
Vol.37 No.19

计 算 机 工 程
Computer Engineering

2011年10月
October 2011

• 软件技术与数据库 •

文章编号：1000—3428(2011)19—0028—04 文献标识码：A 中图分类号：N945

一种基于新型关系矩阵的数据填补方法

金成美[1]，鄂 旭[1,2,3]，穆海军[1]，李 岩[4]

(1. 辽宁工业大学电子与信息工程学院，辽宁 锦州 121001；2. 辽宁工程技术大学资源与环境学院，辽宁 阜新 123000；
3. 辽宁工程职业技术学院，辽宁 铁岭 112000；4. 中国铁通锦州分公司，辽宁 锦州 121000)

摘 要：研究不完备信息系统，分析容差关系、非对称相似关系、限制容差关系的局限性，提出一种基于新型关系矩阵的数据填补方法。新型关系矩阵完整地记录了各对象之间条件属性以及决策属性的异同情况，以此挖掘对象间的潜在联系，并进行空缺值的填补处理，填补的结果不会破坏系统的协调性。数据集测试结果验证了该方法的有效性。

关键词：新型关系矩阵；不完备信息表；粗糙集；数据填补；冲突避免

Data Filling Method Based on New Relationship Matrix

JIN Cheng-mei[1], E Xu[1,2,3], MU Hai-jun[1], LI Yan[4]

(1. School of Electronics & Information Engineering, Liaoning University of Technology, Jinzhou 121001, China;
2. School of Resources and Environment Engineering, Liaoning Technical University, Fuxin 123000, China;
3. Liaoning Engineering and Professional Technology Institute, Tieling 112000, China;
4. China Tietong Jinzhou Branch, Jinzhou 121000, China)

【Abstract】 Under incomplete information system, there are several similarity relations, such as tolerance relation, dissymmetrical similarity relation, limited tolerance relation. But all have limitations respectively. In this paper, a method for completing data based on new relationship matrix is presented. The new relationship matrix records all the situations that similarities or differences fort comparing the condition attributes and the decision attributes between objects. On basis of it, mines the potential links between objects, and completes the missing data. Results will not undermine the system's coordination. Experimental results indicate the method is effective.

【Key words】 new relationship matrix; incomplete information table; rough set; data filling; collision avoidance

DOI: 10.3969/j.issn.1000-3428.2011.19.008

1 概述

粗糙集理论[1-3]是一种处理含糊及非精确性问题的数学工具，在机器学习、决策分析、归纳推理、模式识别、数据挖掘等诸多领域得到广泛应用。但基于传统等价关系的粗糙集理论是不能处理不完备信息系统的。而在现实生活中，由于数据的测量误差、数据理解或获取的限制等原因，使得在知识获取时往往面对的是不完备信息系统。

为使粗糙集理论有效地应用于不完备信息系统，目前已有很多研究和改进，主要方法之一就是对传统的粗糙集理论模型在不完备信息系统中进行扩充[4-8]，再在其基础上完成缺失数据的补齐处理。现在常用的扩充模型是将严格的等价关系放宽为要求更为宽松的容差关系、非对称相似关系、限制容差关系等。应用最为广泛的是容差关系，它对不完备信息系统中的样本对象间的相似性给出了定义，但是容差关系的相似条件过于宽松，非对称相似关系划分相似类的条件又过于苛刻，不利于处理大型数据。例如对象 $x=(1,*,3,4,5,6)$ 与对象 $y=(*,2,3,4,5,6)$ 根据非对称相似关系定义，被划分到不同类中，但实际上 x,y 相等的可能性很大。限制容差关系优于以上2种扩充方法，也更符合实际情况，但是存在以下局限性：个体对象之间只要有一个属性值不同，那么就认为是完全不同的，必须划分到不同的容差类中。这种不允许同一类的个体之间存在丝毫偏差的模型，会导致容差类的划分过细，个数太多，给不完备信息系统的处理带来极大的不便，不适合大型不完备信息系统的处理。

上述扩充模型或是它们的变体，其根本目的都是为区分2个样本提供依据。事实上，无论是非对称相似关系、限制容差关系还是其他变体，都可以看做是在容差关系的基础上对相似性的不同程度的刻画。为更好地避免决策冲突，较好地保持信息表的决策规则，本文从样本对象之间的相异性入手，以粗糙集、扫描向量[4]为理论基础，对可辨识矩阵进行改进，提出一种基于新型关系矩阵的数据填补方法。

2 相关概念及定理

粗糙集理论中包含很多概念[2-3]，这里只介绍与本算法相关的概念。

定义1 设决策表为 $S=(U,A,V,f)$，其中，$U=\{x_1,x_2,\cdots,x_n\}$ 是一个非空有限对象集合；A 是对象的属性集合，分别由条件属性集 $C=\{a_i|i=1,2,\cdots,m\}$ 和决策属性集 $D=\{d\}$ 2个不相交的子集组成，即 $A=C\cup D$；a_i 是样本 x_j 在属性 $a_i(x_j)$ 上的

基金项目：国家自然科学基金资助项目(70971059, 70771007)；辽宁省博士科研启动基金资助项目“基于数据挖掘技术的电信客户市场细分系统研究”(20091034)

作者简介：金成美(1985－)，女，硕士研究生，主研方向：知识发现，智能决策支持系统；鄂 旭，教授、博士；穆海军，硕士研究生；李 岩，工程师

收稿日期：2011-04-20 **E-mail：**jcm0624@hotmail.com

取值。定义新型关系矩阵 $\boldsymbol{VM}(i,j)=(D,C_D,\boldsymbol{F})$，其中，决策标识 D 表示为：

$$D=\begin{cases}1 & d(x_i)\neq d(x_j)\\ 0 & d(x_i)=d(x_j)\end{cases} \tag{1}$$

对象之间的差别属性集 C_D 表示为：

$$C_D=\{a_k\,|\,a_k\in A\wedge a_k(x_i)\neq a_k(x_j)\wedge a_k(x_i)\neq *\wedge a_k(x_j)\neq *\} \tag{2}$$

条件属性标识向量 $\boldsymbol{F}$ 表示为：

$$\boldsymbol{F}=(f(a_1),f(a_2),\cdots,f(a_n)) \tag{3}$$

其中：

$$f(a_i)=\begin{cases}|C_D|^{-1} & a_i\in C_D\\ * & a_i(x_i)=*\vee a_i(x_j)=*\\ 0 & a_i\notin C_D\end{cases} \tag{4}$$

显然，关系矩阵 $\boldsymbol{VM}$ 是对称的、基于向量的矩阵。矩阵中每个元素较完整的记录了相应的样本对象的关系：当 $D=1\wedge C_D\neq\varnothing$，表示两样本对象是异决策有差别的；当 $D=1\wedge C_D=\varnothing\wedge\forall f(a_i)=0$，表示两样本对象是冲突的；当 $D=1\wedge C_D=\varnothing\wedge\exists f(a_i)=*$，表示两样本对象是异决策无差别的；当 $D=0\wedge C_D\neq\varnothing$，表示两样本对象是同决策有差别的；当 $D=0\wedge C_D=\varnothing$，表示两样本对象是同决策无差别的。

在填充缺失数据的过程中，考虑经济因素时需要注意重要属性和冗余属性对决策的贡献不同，所以在进行数据填补时，可根据属性重要度择优填补。本文将给出初步计算核属性的方法，相关定理如下：

定理 1 给定不完备信息表 $S=(U,A,V,f)$ 及其完备的子信息表 S'，其中，$A=C\cup D$，$a_i\in C$，如果 a_i 是 S' 中的核属性，则 a_i 一定也是 S 中的核属性。

证明：设 a_i 是 S' 的核属性，则去除 a_i 后，$\exists x,y\in U'$，$\forall a_j\in C$ 有 $a_j(x)=a_j(y)\wedge d(x)\neq d(y)$，其中，$d\in D$，$S'$ 中产生决策矛盾，又 $\because U'\in U$，去除 a_i 后，S 中同样会产生决策矛盾。因此，可以证明 a_i 同样是 S 中的核属性。

定理 2 在任一个 $\boldsymbol{VM}(i,j)=(D,C_D,\boldsymbol{F})$ 中，如果 $\boldsymbol{VM}(i,j)=(D=1,C_D\neq\varnothing,\forall f(a_i)\neq *\wedge\exists f(a_j)=1)$，则该元素对应的属性 a_j 为原信息表的核属性。

证明：$\because\forall f(a_i)\neq *$，说明 $\forall\boldsymbol{VM}(i,j)$ 所对应的对象 x_i,x_j 是完备的，$x_i,x_j\in S'$，S' 是原信息表的完备子信息表。

$\because\exists f(a_j)=1\wedge f(a_j)=|C_D|^{-1}\Rightarrow|C_D|=1$，即 x_i,x_j 的差别属性集 C_D 中只有一个元素，又 $\because D=1$ 即 $d(x_i)\neq d(x_j)$，说明整个属性集中只有一个属性能区分 x_i,x_j。所以，该属性就是完备的子信息表中的核属性。

根据定理 1，可知该属性同样为原信息表中的核属性。

定理 3(冲突避免）决策表 $S=(U,A,V,f)$，其中，$A=C\cup D$，$x_i\in U$，x_i 中有缺失。则对 x_i 进行填补后，只可能与 $\boldsymbol{VM}(i,j)=(D=1,C_D=\varnothing,(\exists f(a_i)=*))$ 项所对应的 x_j 产生矛盾，而不会与 $\boldsymbol{VM}(i,j)=(D=1\text{ or }0,C_D\neq\varnothing)$ 项所对应的 x_j 产生矛盾。

证明：

(1)针对 $\boldsymbol{VM}(i,j)=(D=1,C_D=\varnothing,(\exists f(a_i)=*))$ 项所对应的 x_j，即 x_i 的异决策无差别对象进行证明。

由于 x_i 和 x_j 在填补后可能出现决策矛盾，即 $\forall a_i\in C\wedge a_i(x_i)=a_i(x_j)\rightarrow\exists(d(x_i)\neq d(x_j))$，显然，唯一能满足要求的就是异决策的无差别对象，即 x_i 填补完毕后，只可能与 $\boldsymbol{VM}(i,j)=(D=1,C_D=\varnothing,(\exists f(a_i)=*))$ 对应的 x_j 产生决策矛盾。

(2)针对 $\boldsymbol{VM}(i,j)=(D=1or0,C_D\neq\varnothing)$ 项所对应的 x_j，即 x_i 的有差别对象进行证明。

$\because C_D\neq\varnothing$ 表明 x_i 和 x_j 必然存在至少一个 $\forall a_i\in C$，$a_i(x_i)\neq a_i(x_j)\neq *$；$\because$ 决策矛盾为：$\forall a_i\in C\wedge a_i(x_i)=a_i(x_j)\rightarrow d(x_i)\neq d(x_j)$，$\therefore x_i$ 在填补后不可能与 $\boldsymbol{VM}(i,j)=(D=1or0,C_D\neq\varnothing)$ 所对应的 x_j 产生决策矛盾。

定义 2[3] 设 $DS=(U,A,F,d)$ 是不完备决策信息系统，S^f 是 DS 的一个选择，且 B^f 是 S^f 的最大分布约简集。若 B^f 是 DS 的所有选择中的最小集合且满足：

$$\min_{x\in U}m_{B_f}^f(x)=\max\min_{x\in U}m_{B_g}^g(x) \tag{5}$$

称缩减的完备信息系统 (U,B_f,f,d) 是 DS 的最优完备选择。

对于不协调不完备决策信息系统，获取决策规则的基本思想是最优选择。最优选择方法排除了人们的主观性，不是通过人主观判断去填补那些缺失的数据，而是找出那些是决策最可能发生的数据。它不是孤立的条件选择，而是通过已知的条件属性值，系统的选择缺失的数据值，即在整体上选择缺失的数据值[3]。不完备数据填补与约简不同，关注的重点并不是冗余属性的消除问题，而是重视丢失数据的合理填补问题。只有合理的填补才能够保证并提高知识约简后规则提取的准确性和可靠性。

3 基于新型关系矩阵的数据填补算法

设新型关系矩阵为 $\boldsymbol{VM}$，并初始化；核属性集 $H=\varnothing$。

Step1 根据原信息表建立 $\boldsymbol{VM}$ 矩阵，同时求出决策系统的缺失对象集 MOS，及遗失对象 x_i 的遗失属性集 $MAS(x_i)$，要求集合 MOS 按$|MAS(x_i)|$由小到大顺序排列。

Step2 搜索 $\boldsymbol{VM}$ 矩阵

if VM(i,j)=(D=1∧C_D={}∧f(a_i)=0)即两样本对象是冲突的，去掉 VM 中的 i 行，i 列，j 行，j 列;

else{

if (VM(i,j)=(D=1,C_D❶{},f(a_i)❶*∧f(a_k)=1))，则修改核属性集，令 H=H⑨{a_k};

}

Step3 如果 MOS 不为空，则转到 Step5；否则，转到 Step4。

Step4 输出 **VM'**及完备的信息表 S'(算法结束)。

Step5 逐一对 MOS 中的缺失对象 x_i 进行如下操作：

if (a_k❺MAS(x_i)∧a_k❺H)

则 a_i 为冗余属性，可任意填补;

else 操作如下:

for (j=1;j<=n;j++) {

If VM(i,j)❶(1,{},(⋯,f(a_k)=*,⋯))

返回原信息表 S，记录 $a_k(x_j)$的值到集合 possiblegetV 中;

else { 返回原信息表 S

if $a_k(x_j)$==*，则放弃;

else 记录 $a_k(x_j)$的值到集合 impossiblegetV 中;}

用集合 possiblegetV 中任意元素填补 $a_k(x_i)$;

令 MAS=MAS\{ a_k}

}

直至 MAS(x_i)为空，修改 VM 的 i 行 i 列，令 MOS= MOS\{x_i}；

Step 6 转到 Step3。

需要说明的是，根据定理 3 对象 x_i 进行填补后，只可能与 $\boldsymbol{VM}(i,j)=\left(D=1,C_D=\varnothing,\left(\exists f(a_i)=*\right)\right)$ 项所对应的 x_j 产生矛盾，而不会与 $\boldsymbol{VM}(i,j)=(D=1\text{ or }0,C_D\neq\varnothing)$ 项所对应的 x_j 产生矛盾。由此，本文提出基于新型关系矩阵的算法，Step5 中利用对象之间的差别关系进行填补，排除了填补过程中产生决策矛盾的可能。如果可能取得的值只有一个，那么就可以直接用这个值进行填补。本文利用差别关系提出的算法，避免上下近似等步骤，降低计算的复杂性。

4 实例研究

4.1 算法应用实例

为考察算法的有效性，选择一个给定不协调的不完备的决策信息表(见表 1)进行对比分析。在表 1 中，a,b,c 为条件属性，d 为决策属性。

表 1 原始决策表

U	a	b	c	d
x_1	1	1	1	1
x_2	1	*	1	1
x_3	2	1	1	1
x_4	1	2	*	1
x_5	1	*	1	2
x_6	2	1	2	2
x_7	1	2	1	2

Step1 把原始信息表转换成关系矩阵：

$$\begin{pmatrix}
(0,\varnothing,(0,0,0)) & (0,\varnothing,(0,*,0)) & (0,\{a\},(1,0,0)) & (0,\{b\},(0,1,*)) & (1,\varnothing,(0,*,0)) & \left(1,\{a,c\},\left(\tfrac{1}{2},0,\tfrac{1}{2}\right)\right) & (1,\{b\},(0,1,0)) \\
 & (0,\varnothing,(0,0,0)) & (0,\{a\},(1,*,0)) & (0,\varnothing,(0,*,*)) & (0,\varnothing,(0,*,0)) & \left(1,\{a,c\},\left(\tfrac{1}{2},*,\tfrac{1}{2}\right)\right) & (1,\varnothing,(0,*,0)) \\
 & & (0,\varnothing,(0,0,0)) & \left(0,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},0\right)\right) & (1,\{a\},(1,*,0)) & (1,\{c\},(0,0,1)) & \left(1,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},0\right)\right) \\
 & & & (0,\varnothing,(0,0,0)) & (1,\varnothing,(0,*,*)) & \left(1,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},*\right)\right) & (1,\varnothing,(0,0,*)) \\
 & & & & (0,\varnothing,(0,0,0)) & \left(0,\{a,c\},\left(\tfrac{1}{2},*,\tfrac{1}{2}\right)\right) & (0,\varnothing,(0,*,0)) \\
 & & & & & (0,\varnothing,(0,0,0)) & \left(0,\{a,b,c\},\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)\right) \\
 & & & & & & (0,\varnothing,(0,0,0))
\end{pmatrix} \tag{6}$$

遗失对象集为：$MOS=\{x_2,x_4,x_5\}$；遗失属性集分别为：$MAS(x_2)=\{b\}$，$MAS(x_4)=\{c\}$，$MAS(x_5)=\{b\}$。由于各对象遗失属性集的势都为 1，因此 MOS 中元素排列顺序不变，$MOS=\{x_2,x_4,x_5\}$。

Step2 搜索 $\boldsymbol{VM}$ 矩阵，$\boldsymbol{VM}(1,7)=\left(1,\{b\},(0,1,0)\right)$，$\boldsymbol{VM}(3,6)=\left(1,\{c\},(0,0,1)\right)$，则修改核属性集 $H=\{b,c\}$。

Step3 由于 MOS 不为空，填补未结束，跳过 Step4，转到 Step5。

Step4 (跳过，下文给出原因)，这里先以对象 x_2 为例。

Step5 因为 $MAS(x_2)=\{b\}$，$b\in H$，所以对矩阵进行搜索。

对于 $\boldsymbol{VM}(i,j)\neq\left(1,\varnothing,\left(\cdots,f(b)=*,\cdots\right)\right)$ 的项，返回原信息表 S，记录 $b(x_j)$ 的值到集合 possiblegetV 中，最终得到 possoblegetV={1}。

将不满足 $\boldsymbol{VM}(i,j)\neq\left(1,\varnothing,\left(\cdots,f(b)=*,\cdots\right)\right)$ 条件的，返回原信息表，记录 $b(x_j)$ 的值到集合 impossiblegetV 中，最终得到 impossiblegetV={2}。由于 possiblegetV 集合中只含有一个元素，所以直接用它进行缺失项的填补。当填补完成后，修改后的 MAS 为空，证明经填补后对象 x_2 为完备的，所以修改关系矩阵 $\boldsymbol{VM}$，得到式(7)。

$$\begin{pmatrix}
(0,\varnothing,(0,0,0)) & (0,\varnothing,(0,0,0)) & (0,\{a\},(1,0,0)) & (0,\{b\},(0,1,*)) & (1,\varnothing,(0,*,0)) & \left(1,\{a,c\},\left(\tfrac{1}{2},0,\tfrac{1}{2}\right)\right) & (1,\{b\},(0,1,0)) \\
 & (0,\varnothing,(0,0,0)) & (0,\{a\},(1,0,0)) & (0,\{b\},(0,1,*)) & (1,\varnothing,(0,*,0)) & \left(1,\{a,c\},\left(\tfrac{1}{2},0,\tfrac{1}{2}\right)\right) & (1,\{b\},(0,1,0)) \\
 & & (0,\varnothing,(0,0,0)) & \left(0,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},0\right)\right) & (1,\{a\},(1,*,0)) & (1,\{c\},(0,0,1)) & \left(1,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},0\right)\right) \\
 & & & (0,\varnothing,(0,0,0)) & (1,\varnothing,(0,*,*)) & \left(1,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},*\right)\right) & (1,\varnothing,(0,0,*)) \\
 & & & & (0,\varnothing,(0,0,0)) & \left(0,\{a,c\},\left(\tfrac{1}{2},*,\tfrac{1}{2}\right)\right) & (0,\varnothing,(0,*,0)) \\
 & & & & & (0,\varnothing,(0,0,0)) & \left(0,\{a,b,c\},\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)\right) \\
 & & & & & & (0,\varnothing,(0,0,0))
\end{pmatrix} \tag{7}$$

对象 x_4,x_5 操作同上。当 x_5 填补完成后，此时 MOS 为空，转到 Step4。

Step4 得到 $\boldsymbol{VM}'$ 以及 S'，如式(8)和表 2 所示。

算法结束。

$$\begin{pmatrix}
(0,\varnothing,(0,0,0)) & (0,\varnothing,(0,0,0)) & (0,\{a\},(1,0,0)) & \left(0,\{b,c\},\left(0,\tfrac{1}{2},\tfrac{1}{2}\right)\right) & (1,\{b\},(0,1,0)) & \left(1,\{a,c\},\left(\tfrac{1}{2},0,\tfrac{1}{2}\right)\right) & (1,\{b\},(0,1,0)) \\
 & (0,\varnothing,(0,0,0)) & (0,\{a\},(1,0,0)) & \left(0,\{b,c\},\left(0,\tfrac{1}{2},\tfrac{1}{2}\right)\right) & (1,\{b\},(0,1,0)) & \left(1,\{a,c\},\left(\tfrac{1}{2},0,\tfrac{1}{2}\right)\right) & (1,\{b\},(0,1,0)) \\
 & & (0,\varnothing,(0,0,0)) & \left(0,\{a,b,c\},\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)\right) & \left(1,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},0\right)\right) & (1,\{c\},(0,0,1)) & \left(1,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},0\right)\right) \\
 & & & (0,\varnothing,(0,0,0)) & (1,\{c\},(0,0,1)) & \left(1,\{a,b\},\left(\tfrac{1}{2},\tfrac{1}{2},0\right)\right) & (1,\{c\},(0,0,1)) \\
 & & & & (0,\varnothing,(0,0,0)) & \left(0,\{a,b,c\},\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)\right) & (0,\varnothing,(0,0,0)) \\
 & & & & & (0,\varnothing,(0,0,0)) & \left(0,\{a,b,c\},\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)\right) \\
 & & & & & & (0,\varnothing,(0,0,0))
\end{pmatrix} \tag{8}$$

原不完备信息表有 3 处属性缺失，6 种完备化选择。根据定义 2 对 6 种完备化选择进行验证计算可知，原不完备信息表有 2 种最优完备选择，分别是 1, 1, 2 和 1, 2, 2(具体计算这里不再赘述)。明显地，第 1 组完备选择中 x_4, x_7 会产生决策冲突，而运用本文方法所得到的完备解正是最优完备选择的第 2 组，有效地避免了由于填补而产生的冲突。

表 2　完备的信息表

U	a	b	c	d
x_1	1	1	1	1
x_2	1	1	1	1
x_3	2	1	1	1
x_4	1	2	2	1
x_5	1	2	1	2
x_6	2	1	2	2
x_7	1	2	1	2

4.2　实验结果

上述实例验证了本文算法的可行性。为进一步证明其有效性，下面分别用特定值法、概率法、本文算法对 UCI 中的 Hepatitis 数据集和 House 数据集进行填补测试，填补结果的识别率如表 3 所示。

表 3　3 种方法在不同数据集上的测试结果　(%)

识别率数据集	特定值法	概率法	本文算法
Hepatitis	68.9	73.4	77.6
House	86.8	89.0	91.5

测试结果表明，本文算法的识别率优于特定值法和概率法。虽然 Hepatitis 数据集和 House 数据集都有将近 50%的样本缺失，但 House 数据集各样本之间的内部联系比 Hepatitis 数据集紧密，在一定程度上影响着数据填补的识别率。

5　结束语

本文算法利用对象之间的差异关系进行数据处理，不需要计算上下近似，就能降低运算复杂度。时间复杂度依赖于不完备信息系统的训练集能够提供数据及内在规律是否足够。最优情况下时间复杂度为 $O(nm)$, $n=card(U)$, $m=card(MOS)$；最坏情况下为 $O(n^2m)$。由于算法结束后会输出完备的关系矩阵，因此可以直接与基于扫描向量的属性约减算法相结合进行计算，为规则提取提供更丰富的信息。需要说明的是，本文算法适用于离散型空缺值填补，如何处理连续型属性的空值填补将是下一步的研究内容。

参考文献

[1] Pawlak Z. Rough Set[J]. International Journal of Computing and Information Sciences, 1982, 11(1): 341-350.

[2] 王国胤. Rough 集理论与知识获取[M]. 西安: 西安交通大学出版社, 2003.

[3] 张文修, 仇国芳. 基于粗糙集的不确定决策[M]. 北京: 科学出版社, 2005.

[4] 鄂　旭, 高学东, 喻　斌. 基于扫描向量的属性约减方法[J]. 北京科技大学学报, 2006, 28(6): 604-608.

[5] Wang Guoyin. Extension of Rough Set Under Incomplete Information Systems[C]//Proc. of 2002 IEEE Int'l Conf. on Fuzzy Systems. Honolulu, USA: IEEE Press, 2002.

[6] Zhang Qinghua, Wang Guoyin, Hu Jun, et al. Incomplete Information Systems Processing Based on Fuzzy Clustering[C]//Proc. of 2006 IEEE/WIC/ACM Int'l Conf. on Web Intelligence and Intelligent Agent Technology. Washington D. C., USA: IEEE Computer Society, 2006.

[7] 邓耀进, 李仁发. 一种粗糙集理论中量化容差关系的改进[J]. 计算机工程与科学, 2009, 31(10): 105-132.

[8] 杨霁琳, 秦克云, 裴　峥. 不完备信息系统中的不可区分关系[J]. 计算机工程, 2010, 36(13): 4-6.

编辑　陈　文

(上接第 27 页)

报表组件进行透视图开发，生成客户端 SWF 文件；使用 JSP 标签库创建展现层的 Flex 透视图标签。

(3)仪表盘报表：基于 JQuery UI 库，实现 iGoogle 风格的分析面板，多个可视化组件组合在一起，实现综合分析。

分别在 Hive 和 HDW 上对 Bizard 系统进行测试。Hive 运行在 1.83 GHz CPU、1 GB 内存的 PC 机上，通过 Hive 的接口导入了一个实际的常用测试数据集 Foodmart，保存在 Hadoop 的 HDFS 中。Foodmart 已经包含了关于销售分析主题的星型架构，即一个 86 837 条元组的销售事实表以及商店维、产品维和时间维等 12 个维度。对多个 MDX 进行查询的结果表明，根据 MDX 包含点查询的个数，响应时间为 5 s~20 s。在更大的数据集上，在每节点配置 3.0 GHz dual core CPU、1 GB 内存的 18 个 Hadoop 节点上验证过 Bizard。对 6 千万条记录数据的测试表明预计算时间不到 5 min，点查询平均响应时间为 0.2 s。

由以上测试可以得出，Bizard 能够对大数据进行图表方式的可视化展示和多维分析，但还要进一步提高交互性强的 Ad hoc 查询分析性能，以提供更好的用户体验。

5　结束语

本文设计并实现了基于 B/S 架构多维可视化分析框架 Bizard，通过将 Hive、HDW 数据库的底层接口封装为 XMLA 的 Discover 和 Execute 接口以及采用视图物化等优化技术，从而支持大规模数据的多维分析。目前，Bizard 仅在 Hive、HDW 上探索了可视化多维分析的可行性，而 HBase 提供了较好的实时查询性能，在下一步工作中将实现 HBase 等 NoSQL 数据库的多维数据访问接口和驱动，以支持实时性能高的 NoSQL 数据库。

参考文献

[1] EMC Corporation. Groundbreaking Study Forecasts a Staggering 988 Billion Gigabytes of Digital Information Created in 2010[EB/OL]. (2007-03-06). http://www.emc.com/about/news/press/us/2007/03062007-4932.htm.

[2] Apache Hadoop Org.. Hadoop[EB/OL]. (2011-02-11). http://hadoop.apache.org.

[3] Nosql-database Org.. NOSQL Databases[EB/OL]. (2011-02-10). http://nosql-database.org/.

[4] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters[C]//Proc. of the 6th Symposium on Operating Systems Design and Implementation. San Francisco, USA: [s. n.], 2004.

[5] Soukup T, Davidson I. 可视化数据挖掘: 数据可视化和数据挖掘的技术与工具[M]. 朱建秋, 葵伟杰, 译. 北京: 电子工业出版社, 2004.

[6] 游进国, 奚建清, 肖裕洪. 基于 PC 集群的并行数据仓库架构[J]. 计算机工程, 2009, 35(20): 73-75.

编辑　任吉慧