粗糙关系数据库的度量

安秋生

(山西师范大学数计学院 临汾 041004)

摘 要 对粗糙关系数据库(Rough Relational Database,RRDB)的度量问题在国内外的发展状况进行了探讨分析。具体分为两部分,其一,对与粗糙关系数据库相关的几个概念进行了介绍;其二,对目前粗糙关系数据库的度量问题在国内外的发展状况进行了总结研究。

关键词 RRDB,近似度量,熵度量,位模式

中图法分类号 TP18 **文献标识码** A

Study of Measures Problem for Rough Relational Database

AN Qiu-sheng

(Mathematics and Computer Science School,Shanxi Normal University,Linfen 041004,China)

Abstract The study of measures problem for rough relational database (RRDB) and its development status were discussed. Concretely, some basic concepts related to rough relational database were given firstly, moreover the development status about measures problem of rough relational database were reviewed and analyzed at home and abroad.

Keywords RRDB, Approximate measures, Entropy measures, Bit pattern

美国学者 Theresa Beaubouef 1993 年把粗糙集与关系数据库结合起来提出了粗糙关系数据库模型(Rough Relational Database Model,RRDM)[1],该模型是处理含糊和不确定数据的一种重要的数据模型,该模型的提出对于智能数据的检索和不确定性问题的解决具有重要的意义。自 RRDM 提出后,国内外学者对其进行了广泛研究,总的说来,其研究是围绕着粗糙关系运算、粗糙函数依赖、RRDB 的范式理论、粗糙数据查询、数据存储与更新及粗糙度量等几个方面进行的。本文首先就与度量相关的几个概念进行讨论,然后就 RRDB 度量问题在国内外的发展进行回顾。

1 RRDB 的相关概念

首先,给出粗糙关系数据库模型及粗糙关系数据库的概念。

定义 1[1] 粗糙关系数据库模型同普通的关系模型一样均是由包含元组的关系构成,元组 t_i 采用$(d_{i1},d_{i2},\cdots,d_{im})$的形式,其中 d_{ij} 是元组的一个属性值,隶属属性域 D_j。在普通的关系数据库模型中,$d_{ij}\subseteq D_j$,而在 RRDM 中,$d_{ij}\subseteq D_j$,$d_{ij}\neq\phi$,一般用 $P(D_i)$ 表示 $D_i-\phi$ 的幂集,D_i 为某一属性域。这里,RRDM 与普通的关系数据库的主要不同点是它的属性值可以由多个原子值构成,而关系数据库则是由单属性值组成。

定义 2[2](RRDB) 我们把 RRDB 定义为三元组:$S=(U,A,D_i)$,从数据库的角度看,U 是所有元组(记录)的集合,A 为数据库的属性集,D_i 为某一属性值域,对于每个 $r\in U$,$a\in A$,有 $r(a)\subseteq D_a$,$r(a)$为 r 在属性 a 上的取值。按照信息系统的定义,RRDB 实际上是特殊的多值信息系统,其中:$U=\{u_1,u_2,\cdots,u_{|U|}\}$是对象的全体,它是非空有限集,即论域;$A=\{a_1,a_2,\cdots,a_{|A|}\}$是属性的全体,它是非空有限集。

在 Theresa Beaubouef 的博士论文中[1],给出了粗糙函数依赖的原始定义。

定义 3 设 R 为粗糙关系数据库模式,T 为其任意粗糙关系,设 X、Y 为粗糙关系模式 R 的属性子集,粗糙函数依赖 $(RFD)X\rightarrow Y$ 对于一个粗糙关系模式 R 的所有实例 T 都成立,当满足:

(1)对于任意两个元组 $t,t'\in\underline{RT}$,下式成立:

redundant$(t(X),t'(X))\rightarrow$redundant$(t(Y),t'(Y))$;

(2)对于任意两个元组 $s,s'\in\overline{RT}$,下式成立:

Rough-redundant$(s(X),s'(X))\rightarrow$Rough-redundant$(s(Y),s'(Y))$。

其中,$\underline{RT}$、$\overline{RT}$ 为 Rough 关系模式 R 的所有实例 T 的下近似与上近似。

2 国内外 RRDB 度量问题研究

所谓 RRDB 度量问题,是指对于粗糙关系数据库的分类问题及属性之间的关系问题等方面的度量研究。下面分别介绍与分析国内外的研究情况。

2.1 Theresa Beaubouef 的熵度量模型

在粗糙关系数据库研究领域,首先对粗糙度量问题展开研究的是美国学者 Theresa Beaubouef,他在文献[3]中使用信息论定义了用于粗糙关系数据库度量的粗糙熵、粗糙模式熵及粗糙关系熵等,下面做简要介绍和分析(定义 4—定义 6)。

定义 4 粗糙集合 X 的粗糙熵 $E_r(X)$为:

到稿日期:2012-02-15 返修日期:2012-07-21 本文受国家自然科学基金(70871072)资助。

安秋生(1966—),男,博士,教授,主要研究方向为粒计算与 Rough 集,E-mail:aaqqss@sina.com。

$E_r(X)=-(\rho_R(X))[\sum Q_i\log(P_i)],i=1,\cdots,n$

式中，i 为等价类序数，$\rho_R(X)$ 为集合 X 的粗糙度，c_i 为等价类 X_i 的基数，$P_i=1/c_i$ 表示等价类中值的平均分布概率，Q_i 表示等价类 i 在论域中的分布概率。

粗糙集的粗糙熵的本质是用于度量“具有相同粗糙度的粗糙集”的分类程度（或不确定程度），其划分越细，粗糙熵越小。

定义 5 粗糙关系模式 S 的粗糙模式熵为：

$E_S(S)=-\sum_j[\sum Q_i\log(P_i)],i=1,\cdots,n;j=1,2,\cdots,m$

式中，对于域 j 来说有 n 个等价类、m 个属性，其余同上。

由于粗糙关系数据库的等价关系存在于值域之上，而不是存在于某个属性上，因此粗糙模式熵反映的是粗糙模式中根据值域所划分的等价类的分类程度，分类越细，粗糙模式熵越小，反之亦然。

而一个确定的粗糙关系实例的粗糙关系熵定义如下。

定义 6 模式实例的粗糙关系熵为：

$E_R(R)=-\sum_j D_{P_j}(R)[\Sigma DQ_i\log(DP_i)],i=1,\cdots,n;j=1,2,\cdots,m$

式中，D_{P_j} 表示关系实例的属性 j 对应的域值的粗糙集的数据库粗糙度，DQ_i 表示某数据库关系元组具有类 i 值的概率，DP_i 表示等价类 i 的某个值在数据库关系中的概率。

显然，模式实例的粗糙关系熵是相对于关系实例而言的，它表明的是某个关系实例越粗糙（不确定元组越多或非单值元素越多），其熵越大。

可以看出，Theresa Beaubouef 的熵度量模型作用如下：粗糙熵 $E_r(X)$ 是对具有相同上下近似、划分不同的等价类的度量，而粗糙模式熵是对值域所划分的等价类的分类程度的度量，粗糙关系熵则是对粗糙关系数据库关系实例元组粗糙程度的度量。

2.2 Michinori Nakata 的兼容度模型

在文献[4]中，日本学者 Michinori Nakata 研究了粗糙关系数据的表示，给出了相应的度量模型，简介如下。

假设 c_α 为关系 r 的约束，则具有 c_α 的元组值 $t[A]$ 的兼容性度量为：

$$Com(c_\alpha|t[A])=\begin{cases}1, & \text{if } \alpha\leqslant Com(c|t[A])\\ Com(c|t[A]), & \text{otherwise}\end{cases}$$

式中，$Com(c|t[A]|)=|t[A_i]\cap\alpha|/|t[A_i]|$。

在此基础上，作者进一步定义了元组及关系的满足度。

定义 7（元组满足度） 设 r 为受约束关系，则具有约束 c_α 的元组的满足度定义为：

$$D_{C_\alpha}=\begin{cases}1, & \text{if } t[\mu]\leqslant Com(c_\alpha|t[A])\\ Com(c_\alpha|t[A]), & \text{otherwise}\end{cases}$$

定义 8（关系满足度） 设 r 为受约束关系，则具有约束 c_α 的关系的满足度定义为：

$D_{C_\alpha}(r)=\min_t D_{C_\alpha}(t)$

在研究函数依赖时，作者定义了属性值针对于某个函数依赖的兼容度。

定义 9 属性值 $t_i[A]$ 针对于某个函数依赖 f 的兼容度为：

$Com(f\mid t_i[A])=\min_{j\neq i}Com(f\mid t_i[A],t_j[A])$

$Com(f\mid t_i[A],t_j[A])=\min(1,1-X_{ij}+Y_{ij})$

$X_{ij}=Com(t_i[X]\ EQ\ t_j[X])$

$Y_{ij}=Com(t_i[Y]\ EQ\ t_j[Y])$

$Com(t_i[X]\ EQ\ t_j[X])=|t_i[X]\cap t_j[X]|/|t_i[X]|\times|t_i[X]|$

Michinori Nakata 所讨论的是元组、关系的满足度，并且涉及到属性值（或集）对于函数依赖的兼容度，这对研究函数依赖的度量有一定的启发意义。

2.3 国内学者关于 RRDB 度量问题的研究

近年来，国内学者对于粗糙关系数据库进行了大量的研究，对于相关的粗糙度量问题也展开了研究，下面进行简要的总结。

邱卫根等学者研究了基于 RST 的粗关系数据库的熵，提出了基于粗集的粗关系模式及其实例的信息熵和粗糙熵的概念，作者引用了下述条件熵的定义：

$$H(Y|X)=-\sum_{i=1}^{s}p(X_i)\sum_{j=1}^{t}p(Y_j|X_i)\log P(Y_j|X_i)$$

并将此定义与粗糙熵定义相结合来研究粗函数依赖的定义及判定条件，该方法的提出对于粗糙函数依赖度量的研究有积极的促进作用。

王丹、吴孟达及刘银山等学者在文献[5]中，研究了粗糙关系数据库的连接算子，并定义了完全连接算子、下近似连接算子及上近似连接算子用于粗糙关系模式分解。在该文中，作者定义了 RRDB 的典型程度、完全程度、不完全程度和缺值程度，并定义了粗糙关系数据库的近似查询精度及近似查询质量。

本人近年来对粗糙关系数据库的相关问题进行了研究，并涉足了粗糙关系的属性间的度量问题，下面简要介绍本人的研究情况。

定义 10[6] 给定任意整数 $p,q_1,\cdots,q_p\geqslant 1$，设 $\Pi_C(T)=\{c_1,c_2,\cdots,c_p\}$，这里 Π 为数据库的投影操作，c_i 为属性 C 对应的属性值。$|\Pi_F(T_{C=c_i})|=q_i$ 为属性值为 c_i 时属性 F 所对应的投影数目，f_i 为频率矢量且 $f_{j|i}$ 表示与每个 c_i 值关联的 F 的相关频率矢量。$\sum_{j=1}^{q_i}f_{j|i}=f_i$ 且 $\sum_{i=1}^{p}f_i=1$，粗糙函数依赖的近似度量可以定义为：

$$\Gamma_{p,q_1,\cdots,q_p}(f_1,f_2,\cdots,f_{q_p})=\sum_{i=1}^{p}f_i(1-f_{j|i})$$

下面给出粗糙函数依赖近似度量所满足的基本性质。

1）零公理。粗糙函数依赖的近似度量满足零公理，即当粗糙函数依赖成立时，其近似度量值为零。

证明：当粗糙函数依赖 $X\to Y$ 成立时，其任意的 $f_{j|i}=0$，按照定义 10，$\Gamma_{p,q_1,\cdots,q_p}(f_1,f_2,\cdots,f_{q_p})=\sum_{i=1}^{p}f_i(1-f_{j|i})=\sum_{i=1}^{p}f_i(1-1)=0$。

即粗糙函数依赖的近似度量值为 0，因此，该度量满足零公理。

2）对称公理。可以形式化地表述为：对所有的 $q\geqslant 1$ 且 $1\leqslant g\leqslant k\leqslant q$，则有：

$\Gamma_q([\cdots,f_g,\cdots,f_k,\cdots])=\Gamma_q([\cdots,f_k,\cdots,f_g\cdots])$

它意味着频率矢量在近似度量公式中出现的顺序将不影响最终的度量值。

证明：$\Gamma_{p,q_1,\cdots,q_p}(f_1,f_2,\cdots,f_g,\cdots,f_k,\cdots,f_{q_p})=\sum_{i=1}^{p}f_i(1-$

（下转第 54 页）

参 考 文 献

[1] 马祖长，孙怡宁，梅涛. 无线传感器网络综述[J]. 通信学报，2004，25(4)：114-124

[2] 王行甫，刘志强，黄秋原，等. WSN 中一种改进的边界盒定位算法[J]. 计算机工程，2011，37(20)：57-59

[3] Bulusu N，Hedemann J，Estrn D. GPS-less low cost outdoor localization for very small devices[J]. IEEE Personal Communications，2000，7(5)：28-34

[4] Iculescu D N，Nath B. DV based positioning in ad-hoc networks [J]. Telecommunication System，2003，22(1-4)：267-280

[5] Simic S N，Sastry S. Distributed localization in wireless ad-hoc networks[EB/OL]. (2002-04-10). http://www.eecs.berkeley.edu/Pubs/TechRpts/2002/4010.html

[6] He Tian，Huang Cheng-du，Blum B M. Range-free localization schemes in large scale sensor networks[C]// Proc of the 9th Annual International Conference on Mobile Computing and Networking. San Diego，USA：ACM Press，2003：81-95

[7] Shang Yi，Ruml W，Zhang Ying. Localization from mere connectivity[C]//Proc of the 4th ACM Int'l Symp on Mobile Ad hoc Networking & Computing. New York：ACM Press，2003：201-212

[8] 黄艳，臧传治，于海斌. 基于改进粒子群优化的无线传感器网络定位算法[J]. 控制与决策，2012，27(1)：156-160

[9] 王新芳，张冰，冯友兵. 基于粒子群优化的改进加权质心定位算法[J]. 计算机工程，2012，38(1)：90-95

[10] Kulkarni R V，Venayagamoorthy G K. Particle swarm optimization in wireless-sensor networks：a brief survey[J]. IEEE Transactions on systems，man，and cybernetics，2011，41(2)：262-267

[11] Gopakumar A，Jacob L. Localization in wireless sensor networks using particle swarm optimization[C]// Proc of IET Int. Conference Wireless，Mobile Multimedia Netw.，2008：227-230

[12] Low K S，Nguyen H A，Guo Hao. A particle swarm optimization approach for the localization of a wireless sensor networks[C]// Proc of IEEE Int. Symp. Ind. Electron. 2008：1820-1825

[13] Shi Yu-hui，Eberhart R C. A modified particle swarm optimization[C]// Proc of the IEEE Conference on Evolutionary Computaton. 1998：69-73

[14] Shi Yu-hui，Eberhart R C. Empirical study of particle swarm optimization[C]// Proc of Congress on Evolutionary Computation. Washington：IEEE，1999：1945-1950

[15] 陈贵敏，贾建援，韩琪. 粒子群算法的惯性权值递减策略研究[J]. 西安交通大学学报，2006，40(1)：53-61

[16] Angeline P J. Using selection to improve particle swarm optimization[C]// Proc of IEEE International Conference on Evolutionary Computation. Anchorage，Alaska，USA，1998：84-89

[17] 陈星舟，廖明宏，林建华. 基于粒子群优化的无线传感器网络节点定位改进[J]. 计算机应用，2010，30(7)：1736-1738

[18] 陈志奎，司威. 传感器网络的粒子群优化定位算法[J]. 通信技术，2011，44(1)：102-104

(上接第 15 页)

$f_{j|i})=f_1(1-f_{j|1})+f_2(1-f_{j|2})+\cdots+f_g(1-f_{j|g})+\cdots+f_k(1-f_{j|k})+\cdots+f_p(1-f_{j|p})=f_1(1-f_{j|1})+f_2(1-f_{j|2})+\cdots+f_k(1-f_{j|k})+\cdots+f_g(1-f_{j|g})+\cdots+f_p(1-f_{j|p})=\Gamma_{p,q_1,\cdots,q_p}(f_1,f_2,\cdots,f_k,\cdots,f_g,\cdots,f_{q_p})$。

3)单调公理。形式化地可以描述为：对所有的 $q'\geqslant q\geqslant 2$，$\Gamma_{q'}([1/q',\cdots,1/q'])\geqslant\Gamma_q([1/q,\cdots,1/q])$。粗糙函数依赖的近似度量(定义 10)满足单调公理。

4)权重和公理。对于所有的 $p\geqslant 2$ 及 $q_1,\cdots,q_p\geqslant 1$，$\Gamma_{p,q_1,\cdots,q_p}([f_1,\cdots,f_q],[f_{1|1},\cdots,f_{q_1|1}],\cdots,[f_{1|p},\cdots,f_{q_p|p}]=\sum_{i=1}^{p}f_i\Gamma_{1,q_i}(f_{1|i},\cdots,f_{q_i|i}])$。比较定义 10 与关系数据库的权重和公理，它们的实质是一样的，因此函数依赖的近似度量满足权重和公理。

对于分组公理，由于数据库取值是单值的，而 RRDB 是多值的，其值的分布具有不确定性，因此该公理未必适合粗糙函数依赖。

另外，在近期研究中，本人给出了位模式(二进制模式)下粗糙函数依赖近似度量公式。

定义 11[7] 给定任意整数 $p,q_1,\cdots,q_p\geqslant 1$，设 $\Pi_C(T)=\{Bit_{C_1},Bit_{C_2},\cdots,Bit_{C_p}\}$，这里 Π 为数据库的投影操作，P 为针对属性 C 的属性值的投影总数，Bit_{C_i} 为某个属性值 c_i 的位模式，$|\Pi_C(T_{C=Bit_{Ci}})|=q_i$ 为属性值等于 Bit_{C_i} 时属性 C 所对应的投影数目，f_i 为 Bit_{C_i} 的频率矢量且 $f_{j|i}$ 表示与每个 Bit_{C_i} 值关联的 C 的相关频率矢量。$\sum_{j=1}^{q_i}f_{j|i}=f_i$ 且 $\sum_{i=1}^{p}f_i=1$，粗糙函数依赖的近似度量可以定义为：

$$\Gamma_{p,q_1,\cdots,q_p}(f_1,f_2,\cdots,f_{q_p})=\sum_{i=1}^{p}f_i(1-f_{j|i})$$

$$=\sum_{i=1}^{p}\frac{Card(Bit_{C_i})}{Card(U)}\left(\left(1-\frac{Card(Bit_{D_j})}{Card(Bit_{C_i})}\right)\right)$$

式中，$Card()$ 表示某个位模式二进制位的基数，$Card(Bit_{D_j})$ 为它与 Bit_{C_i} 相关联的属性值位模式的交运算结果不为“0”的值基数加 1(含第一个参与运算的属性值)。

结束语 本文对粗糙关系数据库度量问题进行了回顾与总结。由于本人掌握的文献资料有限，因此难免有这样那样的不足，粗糙关系数据库需要研究与完善的问题很多，度量问题只是其一，希望本文能起到抛砖引玉的作用，引起各位研究者的兴趣，以推动该研究领域的发展。

参 考 文 献

[1] Beaubouef T. Uncertainty processing in a relational database model via a rough set representation[D]. University Microfilms International. A Bell&Howell Information Company，1994：70-72

[2] 安秋生，徐久成，王国胤，等. 基于粗糙关系数据库的粗糙数据查询[J]. 西安交通大学学报，2002，36(8)：859-862

[3] Beaubouef T，Petry F E，Arora G. Information-theoretic measures of uncertainty for rough sets and rough relational databases[J]. Journal of Information Sciences，1998，109：185-195

[4] Nakata M，Murai T. Data Dependencies over Rough Relational Expressions[C]// IEEE Intl. Fuzzy Systems Conf. 2001：1543-1546

[5] 王丹，吴孟达，刘银山. 粗糙关系数据库空间结构及其粗糙集模型[J]. 计算机工程与应用，2005(34)：163-167

[6] 安秋生. 粗糙函数依赖的近似度量[J]. 计算机工程与应用，2009，45(1)：144-146

[7] 安秋生. 位模式下粗糙函数依赖近似度量的研究[J]. 计算机工程与应用，2011，47(2)：26-28